Trusted Artificial Intelligence in Manufacturing

A Review of the Emerging Wave of Ethical and Human Centric AI Technologies for Smart Production

John Soldatos and Dimosthenis Kyriazis
(Editors)

now
the essence of knowledge

Published, sold and distributed by:
now Publishers Inc.
PO Box 1024
Hanover, MA 02339
United States
Tel. +1-781-985-4510
www.nowpublishers.com
sales@nowpublishers.com

Outside North America:
now Publishers Inc.
PO Box 179
2600 AD Delft
The Netherlands
Tel. +31-6-51115274

ISBN: 978-1-68083-876-3
E-ISBN: 978-1-68083-877-0
DOI: 10.1561/9781680838770

Suggested citation: John Soldatos and Dimosthenis Kyriazis (eds.). (2021). *Trusted Artificial Intelligence in Manufacturing: A Review of the Emerging Wave of Ethical and Human Centric AI Technologies for Smart Production*. Boston–Delft: Now Publishers

Disclaimer: The various chapters of the book reflect only the authors' views. The European Commission is not responsible for any use that may be made of the information contained in the book.

This book has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 956573.

Table of Contents

Preface

Nowadays, industrial organizations are heavily investing in the digital transformation of their production processes as part of their transition to the fourth industrial revolution (Industry4.0). Based on Cyber Physical Systems (CPS) and backbone technologies like cloud computing, Industrial Internet of Things (IIoT) and Artificial Intelligence (AI), Industry4.0 is contributing towards the realization of flexible production lines, while supporting innovative functionalities like mass customization, predictive maintenance, Zero Defect Manufacturing (ZDM), and digital twins. AI is currently the most disruptive digital enabler of the Industry4.0 era. It facilitates novel use cases like predictive quality management (Quality 4.0), effective human-robot interaction and collaboration, agile production and generative product design. AI's disruptive potential is propelled by advances in hardware and scalable software systems, which have allowed the efficient utilization of advanced machine learning frameworks and novel algorithms that are suitable for large scale problems in realistic settings. Thus, advanced AI technologies tackle challenges ranging from large scale optimization to control problems in manufacturing environments.

Despite these advances, state of the art AI deployments in manufacturing do not take full advantage of the latest innovative capabilities of machine and deep learning, as well as of robotic systems. Rather, they are sophisticated to a limited extent and are mostly focused on the consolidation of datasets from heterogeneous sources towards enabling advanced analytics (e.g., deep learning) for use cases such as predictive maintenance and industrial simulations. Real-life manufacturing environments are complex, dynamic and unpredictable, which highlights safety, reliability and trustworthiness challenges for the respective AI deployments. Specifically, real-life deployments of advanced AI systems face challenges in the following areas:

(A) Transparency and Explainability: Nowadays data scientists can understand the algorithms used to train AI systems. However, they are in most cases unable to reason about their functionality and to explain their operation. As a result, the operation of AI systems is usually not transparent and therefore human users cannot fully understand their behaviour. Likewise, manufacturing employees are reluctant to trust AI solutions and accept their deployment in the shopfloor.

(B) AI systems Interaction with the Manufacturing Environment: The successful deployment of AI systems in the manufacturing shopfloor requires their effective interaction with the surrounding environment, including cyber and physical systems. Such interactions are fundamental for the success of certain machine learning models, such as Reinforcement Learning (RL). Nevertheless, the interaction between AI systems and other applications is still typically slow (e.g., non real-time), hazardous and risky (e.g., prone to mistakes that could cause physical damage). This is a serious barrier for deploying AI systems at scale i.e., systems that involve many interactions between AI systems and other elements of the surrounding environment (e.g., software and physical systems).

(C) Human Centric AI Systems: Despite the expanded use of AI in factories, employees remain the most flexible resource. In the years to come, employees, AI solutions and robots will co-exist in the manufacturing environment. Thus, AI systems must be human centric i.e., able to consider the context of the employee and dynamically adapt to it. Likewise, employees must be properly trained to use and co-exist with AI solutions in the workplace. This human centred operation of AI systems can be very challenging and is not adequately addressed by state-of-the-art digital manufacturing platforms. This is also one of the reasons why many AI systems operate in isolation from humans.

(D) AI Cybersecurity Challenges: The deployment of AI systems in the shopfloor raises significant security challenges. For example, it makes it possible for attackers to compromise the operation of a deep neural network either through taking control over the system or through altering its input data (i.e. poisoning) in a way that outputs malicious decisions. As another example, an adversary can attack an AI system towards accessing confidential data or proprietary learning models that could lead to IP (Intellectual Property) theft.

(E) Inaccuracy and Unreliability of Industrial Data: The quality and the quantity of the data that are used for building AI systems are a decisive factor for their proper functioning. For example, limited data quality can be a source of poorly performing or biased AI systems. Unfortunately, industrial data are inherently unreliable, as readings from CPS systems and IoT devices can be skewed by high temperatures, human errors, hardware malfunctions or even cyber-attacks. Therefore, data reliability is a major challenge when building AI systems for use cases like quality management, agile production operations and human robot collaboration.

Overall, the successful deployment of AI solutions in manufacturing environments hinges on their security, safety and reliability which becomes more challenging in settings where multiple AI systems (e.g., industrial robots, robotic cells, Deep Neural Networks (DNNs)) interact with humans. The safe, reliable and trustworthy operation of AI systems at scale is a key perquisite for establishing confidence in their behaviour and operation. To guarantee the safe and reliable operation of AI systems in the shopfloor, there is a need to address many challenges in the scope of complex, heterogeneous, dynamic and unpredictable environments. Specifically, data reliability, human machine interaction, security, transparency and explainability challenges need to be addressed at the same time. Recent advances in AI research (e.g., in deep neural networks security and explainable AI (XAI) systems), coupled with novel research outcomes in the formal specification and verification of AI systems provide a sound basis for safe and reliable AI deployments in production lines. However, the legal and regulatory dimension of safe and reliable AI solutions in production lines must be considered as well. Hence, the development of technical solutions for the robust, secure and safe operation of AI systems in manufacturing, along with the study of the legal implications of safe and secure AI in production lines are key prerequisites towards an ethical AI in manufacturing as illustrated in the guidelines of EU's High Level Expert Group (HLEG) on AI and reflected in the emerging EU regulation for AI.

To address some of the above listed challenges, fifteen European Organizations collaborate in the scope of the STAR project, a research initiative funded by the European Commission in the scope of its H2020 program (Grant Agreement Number: 956573). Specifically, STAR is a joint effort of AI and digital manufacturing experts towards enabling the deployment of standard-based secure, safe reliable and trusted human centric AI systems in real-life manufacturing environments. STAR researches, develops, and validates novel technologies that enable AI systems to acquire knowledge in order to take timely and safe decisions in dynamic and unpredictable environments. Moreover, the project researches and will deliver approaches that enable AI systems to confront sophisticated adversaries and to remain robust against security attacks. In this way STAR's solutions eliminate security and safety barriers that hinder the deployment of sophisticated AI systems in real-life production lines. STAR will produce technical solutions that boost the safety, robustness and trustworthiness of systems AI in dynamic, real-life settings, while at the same exploring the legal implications of a safe and secure AI in prominent manufacturing scenarios.

This book is co-authored by the STAR consortium partners and aims at providing a complete and comprehensive review of technologies, techniques and systems for trusted, ethical, and secure AI in manufacturing. The different chapters of the book cover systems and technologies for industrial data reliability, responsible and transparent artificial intelligence systems, human centered manufacturing systems such as human centred digital twins, cyber-defence in AI systems, simulated reality systems, human robot collaboration systems, as well as automated mobile robots for manufacturing environments. A variety of cutting-edge AI technologies are employed by these systems including deep neural networks, reinforcement learning systems, and explainable artificial intelligence systems. Furthermore, relevant standards and applicable regulations are discussed.

Beyond reviewing state of the art standards and technologies, the book illustrates how the STAR research goes beyond the state of the art, towards enabling and showcasing human centred technologies in production lines. Emphasis is put on dynamic human in the loop scenarios, where ethical, transparent and trusted AI systems co-exist with human workers.

The book consists of 11 Chapters:

- **Chapter 1 ("Blockchain Based Data Provenance for Trusted Artificial Intelligence")** deals with blockchain based solutions for industrial data reliability. It presents the advantages of blockchain technologies for tracking and tracing industrial data. It also reviews different blockchain solutions for digital manufacturing, including data provenance and reliability solutions. Additionally, the chapter outlines a solution for tracing data and metadata of AI algorithms for industrial applications.
- **Chapter 2 ("Artificial Intelligence and Secure Manufacturing: Filling Gaps in Making Industrial Environments Safer")** presents cybersecurity solutions for AI systems in industrial settings. Specifically, it analyses the security challenges of AI solutions for smart manufacturing environments. The analysis focuses on the adversarial models utilized by malevolent entities to cause malfunctions to AI-powered systems. Moreover, the chapter presents state-of-the-art approaches for securing machine-learning models, including deep neural networks. Emphasis is put on attestation-based provenance mechanisms that guarantee the trustworthiness of data streams feeding AI systems. Likewise, robust solutions that mitigate adversarial machine learning attacks are also introduced.
- **Chapter 3 ("Knowledge Modelling and Active Learning in Manufacturing")** is devoted to active learning solutions for manufacturing environments. It illustrates how the use of active learning techniques for identifying the most informative data instances for which to obtain users' feedback, reduce

friction, and maximize knowledge acquisition. Moreover, the chapter presents the merits of combining semantic technologies and active learning in manufacturing use cases.

- **Chapter 4 ("Multimodal Human Machine Interactions in Industrial Environments")** reviews Human Machine Interaction (HMI) techniques for industrial applications, with emphasis on multimodal interactions between industrial machines and robots. Furthermore, the chapter provides examples and use cases in fields related to multimodal interaction in manufacturing, such as augmented reality. The chapter concludes by discussing the deployment and use of AI and multimodal HMI in the context of the various applications in production lines.
- **Chapter 5 ("A Review of Explainable Artificial Intelligence in Manufacturing")** provides an overview of Explainable Artificial Intelligence (XAI) techniques in manufacturing applications. It presents how XAI can boost the transparency of AI models and analyses different metrics that can used to evaluate XAI techniques. Moreover, the chapter illustrates practical applications of XAI techniques in the manufacturing domain.
- **Chapter 6 ("Confidence Assessment of AI Models in Simulated Industrial Environments")** discusses the importance of artificially generated adversarial scenarios for assessing an AI agent's confidence level and quality. It also presents techniques that aim to increase the confidence assessment of manufacturing focused AI agents, including techniques that span the fields of Reinforcement Learning, Explainable AI and Visual Analytics.
- **Chapter 7 ("The Human-Digital Twin in the Manufacturing Industry: Current Perspectives and a Glimpse of Future")** explains why and how manufacturing workers must nowadays interact with complex production systems under challenging conditions. Accordingly, it illustrates how human centric Digital Twins can alleviate these challenges. The chapter introduces an anatomy of human centred digital twins that can represent humans in the digital world, including their intents, behaviours, conditions, and emotions. It also explains how such digital twins provide the ground for human-aware operations and planning.
- **Chapter 8 ("Video Analytics for Situation Awareness Safe Robot-Human Cohabitation in Production Lines")** focuses on solutions for dynamically detecting safety zones in human robot collaboration scenarios. Specifically, it presents algorithms that analyse scenes from the plant using the global point of view of a camera network deployed in the factory. Video analytics are used to detect anomalies and to raise alarms in a timely fashion. Emphasis is put on the presentation of techniques that detect objects of interest in video streams

and localize them in the 3D environment. The purpose of these video analytics is to feed a "planner" indicating dynamically the areas that should be avoided by a robots' fleet operating in the production lines.

- **Chapter 9 ("Human in the Loop of AI Systems in Manufacturing")** reviews systems and technologies that empowers humans and AI actors to work in synergy. Moreover, the chapter considers the potential emergent outcomes of such a synergy in a way that goes beyond automation or augmentation. A model of human-AI interaction is presented, along with techniques for increasing the efficiency of human-AI collaboration.
- **Chapter 10 ("A Review of Industrial Standards for AI in Manufacturing")** provides a review of industrial standards related to AI solutions in manufacturing environments, including: (i) Recommendations for human centric manufacturing systems; and (ii) Technical standards for safety, security and data management.
- **Chapter 11 ("AI That Works")** covers AI-related organizational and management issues, beyond AI technologies. It presents notions and guidance to make AI work rather than just function. The chapter promotes an "AI that works by design" disciplines that prepares AI to work by design with embedded non-functionals for cases when things may go wrong and other risks it may encounter or cause.

The book is made available as an open access publication, which could make it broadly and freely available to the AI and smart manufacturing communities. We would like to thank "now publishers" for the opportunity and their collaboration in making this happen. Most importantly, we take the chance to thank all contributing authors for their valuable inputs and collaboration. Finally, we would also like to acknowledge funding and support from the European Commission as part of the H2020 STAR project, which made this open access publication possible.

June 2021
John Soldatos
Dimosthenis Kyriazis

Glossary

A

AGV - *Automated Guided Vehicle*. 140

AI - *Artificial Intelligence*. 1–3, 5, 11–15, 20–22, 24, 30–32, 34, 37, 40, 42–48, 74, 76, 87, 93–95, 98, 100, 106–108, 114, 115, 120, 122–124, 126, 173–176, 178, 179, 181, 185–187

AK - *Attestation Key*. 44

AM - *Additive Manufacturing*. 5, 8

AMQP - *Advanced Message Queue Protocol*. 175

AOM - *Analytics Orchestrator Manifest*. 20, 21

APD - *Analytics Processor Definition*. 19, 20

APM - *Analytics Processor Manifest*. 20

AR - *Augmented Reality*. 82, 83

B

B2MML - *Business to Manufacturing Markup Language*. 174

BBA - *Binary-Based Attestation*. 44

BDVA - *Big Data Value Association*. 185, 186

BFO - *Basic Formal Ontology*. 54, 55

C

CAM - *Class Activation Mapping*. 98, 101, 107

CBM - *Constraint-Based Modeling*. 77

CERT - *Computer Emergency Response Teams*. 14

CFA - *Control Flow Attestation*. 44

CM - *Conceptual Model*. 185

CMMS - *Computerized Maintenance Management System*. 2

CNN - *Convolutional Neural Network*. 79–82, 116, 117, 119, 149, 151–154

CPPS - *Cyber Physical Production Systems*. 2, 7, 14, 18, 19, 181

CPS - *Cyber-Physical Systems*. 7, 34, 133

CPU - *Central Processing Unit*. 40

CRM - *Customer Relationship Management*. 175

CRUD - *Create Update Delete*. 4

CSIRT - *Computer Security Incident Response Teams*. 14

CVAE - *Convolutional Variational Auto-Encoder*. 117

CVaR - *Conditional Value at Risk*. 122

D

DI - *Data Interface*. 19

DK - *Data Kind*. 19, 20

DNN - *Deep Neural Network*. 42, 79, 120, 123, 125

DNS - *Domain Name System*. 33

DOLCE - *Descriptive Ontology for Linguistic and Cognitive Engineering*. 54–56

DPM - *Deformable Part Model*. 153

DSM - *Data Source Manifest*. 19–21

DT - *Digital Twin*. 135, 138, 141, 143

DXF - *Drawing Interchange Format*. 174

E

ERP - *Enterprise Resource Planning*. 2, 53, 175, 183

ESCO - *European Skills/Competences, Qualifications and Occupations*. 137

ETSI - *European Telecommunications Standards Institute*. 175, 178

EU - *European Union*. 194, 207

F

FDI - *Factory Design and Improvement*. 174

FSL - *Few-shot Learning*. 116

G

GAAL - *Generative Adversarial Active Learning*. 60

GAN - *Generative Adversarial Network*. 60, 117, 119

GDP - *Gross Domestic Product*. 194

GFO - *General Formal Ontology*. 54, 55

GIGO - *Garbage In Garbage Out*. 3

GMM - *Gaussian Mixture Model*. 151, 155

GNN - *Graph Neural Network*. 123

GradCAM - *Gradient-weighted Class Activation Mapping*. 98, 101

GSR - *Galvanic Skin Response*. 138

H

H-CPS - *Human Cyber-Physical-System*. 134, 135

H2M - *Human-to-machine*. 194, 195

H2M2M - *Human-to-machine-to-machine*. 194, 195

HDT - *Human Digital Twin*. 134–143

HIRL - *Human Intervention Reinforcement Learning*. 122

HMI - *Human Machine Interaction*. 73, 76

HOG - *Histogram of Oriented Gradients*. 153, 154

HR - *Heart-Rate*. 138, 140

HRC - *Human Robot Collaboration*. 80–82

HRV - *Heart Rate Variability*. 140

I

ICT - *Information and Communications Technology*. 32

IDS - *Intrusion Detection System*. 42, 45

IEC - *International Electrotechnical Commission*. 175–179, 182, 185, 187

IGES - *Initial Graphics Exchange Specification*. 174

IIC - *Industrial Internet Consortium*. 184

IIoT - *Industrial Internet of Things*. 31, 181, 183, 184

IOT - *Internet of Things*. 139

IoT - *Internet of Things*. 2, 5, 7, 8, 22, 23, 31, 34, 35, 37–40, 44–47, 115, 178, 183–185

IP - *Intellectual Property*. 6, 8

IRTF - *Internet Research Task Force*. 175

ISMS - *Information Security Management System*. 176, 177

ISO - *International Organization for Standardization*. 174–181, 185–187

ITU-T - *International Telecommunication Union*. 175

L

LIME - *Local Interpretable Model-agnostic Explanations*. 98, 99, 123

LRP - *Layer Wise Relevance Propagation*. 98, 101, 107, 123, 124

LSTM - *Long short-term memory*. 47, 48, 80, 82

M

M2H - *Machine-to-human*. 194, 195

M2M - *Machine to Machine*. 142

MEC - *Mobile Edge Computing*. 38

MES - *Manufacturing Execution Systems*. 53, 175

ML - *Machine Learning*. 2, 11, 12, 20, 41, 47, 115, 116, 120, 123–126, 181, 185

MQTT - *Message Queuing Telemetry Transport*. 139, 142, 176

MR - *Mixed Reality*. 82

MRO - *Manufacturing Reference Ontology*. 54, 55

MSE - *Manufacturing System Engineering*. 55

MSO - *Manufacturing Systems Ontology*. 54, 55

N

NASA - *National American Space Agency*. 135

NIST - *National Institute of Standards and Technology*. 175

NLP - *Natural Language Processing*. 74, 76–78, 87

NST - *Neural Style Transfer*. 117

O

O*NET - *Occupational Information Network*. 137

OAGIS - *Open Applications Group Integration Specification*. 174

OECD - *Organisation for Economic Cooperation and Development*. 194, 195

OEM - *Original Equipment Manufacturer*. 199

OOEF - *OASIS Open Europe Foundation*. 175

OPC-UA - *Open Platform Communications United Architecture*. 138

OT - *Operational Technology*. 183

P

P-PSO - *Politecnico di Milano–Production Systems Ontology*. 54, 55

P-VMDNN - *Predictive Visuo-Motor Deep Dynamic Neural Network* . 81

PBA - *Property-based Attestation*. 44, 45

PCA - *Privacy Certification Authority*. 44

PCR - *Platform Configuration Registers*. 44

PKI - *Public Key Infrastructure*. 45

PLC - *Programmable Logic Controller*. 138, 141, 143, 183

PLM - *Product Lifecycle Management*. 175

PoS - *Proof of Stake*. 14

PoW - *Proof of Work*. 14

PPR - *Product, Process, Resource*. 174

R

R-CNN - *Region Based Convolutional Neural Networks*. 152, 153

RA - *Reference Architecture*. 185

RAMI4.0 - *Reference Architecture Model Industrie 4.0*. 181–183

RCAN - *Randomized-to-Canonical Adaptation Network*. 119

RDC - *Resolver Digital Converter*. 84

ReLU - *Rectified Linear Unit*. 97

RGB - *Red, Green, Blue*. 74, 78, 80, 81

RL - *Reinforcement Learning*. 118, 119, 121, 122

RNN - *Recurrent Neural Network*. 80, 82

ROAR - *RemOve And Retrain*. 103

ROI - *Return-on-Investment*. 2, 35

ROS - *Robot Operating System*. 84–86

RUL - *Remaining Useful Life*. 19

S

S.I.M. - *Scenario, Impact, Measures*. 201, 202

SAI - *Securing Artificial Intelligence*. 178

SC - *Skin Conductance*. 138, 193, 194

SCADA - *Supervisory control and data acquisition*. 2, 40, 183

SCM - *Supply Chain Management*. 175

SDG - *Sustainable Development Goals*. 195, 206, 208

SDG - *United Nations' Sustainable Development Goals*. 193

SDK - *Software Development Kit*. 86, 139

SDO - *Standards Development Organization*. 174, 175

SHAP - *SHapley Additive exPlanations*. 99, 100

SoS - *Systems of Systems*. 31, 39

SSD - *Single Shot Detector*. 152

SSO - *Standards Setting Organization*. 175

STAR - *Safe and Trusted Human Centric Artificial Intelligence in Future Manufacturing Lines*. 149, 150, 154, 155

SUMO - *Suggested Upper Merged Ontology*. 54

SVM - *Support Vector Machines*. 46

T

TCG - *Trusted Computing Group*. 44

TOSCA - *Topology and Orchestration Specification for Cloud Applications*. 176

TPM - *Trusted Platform Module*. 43, 44

V

VA - *Visual Analytics*. 124–126

VGG - *Visual Geometry Group*. 153

VR - *Virtual Reality*. 82–86

VRML - *Virtual Reality Modelling Language*. 174

W

W3C - *World Wide Web Consortium*. 175, 177

X

XAI - *Explainable Artificial Intelligence*. 93–96, 101, 102, 106–108, 115, 123, 124, 126

XKMS - *Xml Key Management Specification*. 177

XML - *eXtensible Markup Language*. 177, 178

XR - *Extended Reality*. 74, 76, 82, 83, 87

Y

YOLO - *You Only Look Once*. 152

DOI: 10.1561/9781680838770.ch1

Chapter 1

Blockchain Based Data Provenance for Trusted Artificial Intelligence

By John Soldatos, Angela-Maria Despotopoulou, Nikos Kefalakis and Babis Ipektsidis

Data reliability is a prerequisite for the development of effective and trusted Artificial Intelligence (AI) systems in industrial environments. Unfortunately, industrial data tend to be unreliable for a variety of reasons (e.g., environmental influence, background noise, and sensor failures). This chapter presents the advantages of blockchain technologies for tracking, tracing, and boosting the reliability of industrial data. It also reviews different blockchain solutions for digital manufacturing, including data provenance and reliability solutions. The chapter ends-up presenting a complete solution for tracing data and metadata of AI algorithms for industrial applications. The solution ensures the use of "sealed" AI algorithms leveraging the properties that render blockchains resilient to tampering. Its main function is to persist the metadata of the algorithms, as well as their outcomes (e.g., prediction

and classification outcomes). As such, it facilitates the implementation strategies that secure AI systems in production lines and boost a trusted AI environment in manufacturing applications.

1.1 Introduction

During the past decade industrial organizations have accelerated their digital transformation and their transition to the fourth industrial revolution (Industry 4.0). This transition is empowered by the deployment of a proliferating number of internet connected systems (e.g., embedded sensors and other Internet of Things (IoT) devices) and of Cyber Physical Production Systems (CPPS) in the manufacturing shopfloor. IoT devices and CPPS systems enable the collection of digital data from physical production processes [1]. Likewise, the analysis of these data drives several optimizations in areas like logistics, quality management, assets maintenance, process control and supply chain management. These optimizations are implemented based on advanced analytics technologies, including BigData analytics, Machine Learning (ML) and Artificial Intelligence (AI). The outcomes of such analytics are used to optimize production processes and to improve manufacturing decisions.

In recent years, Artificial Intelligence (AI) is considered the most disruptive and impactful digital enabler of Industry 4.0. AI systems and algorithms are nowadays enabling manufacturing use cases with a proven Return on Investment (ROI), such as predictive maintenance [2], predictive quality management (Quality 4.0) [3], zero-defect manufacturing [4] and generative product design [5]. The rise of AI in manufacturing use cases is driven by the explosion of available data points, the accelerated increase of available computational capacity, as well as advances in AI software frameworks and tools. Data availability is a key prerequisite for the development of effective AI systems in manufacturing, which is the main reason why most national strategies for AI consider the transition to a data economy as a critical success factor for the AI era [6].

Unfortunately, industrial enterprises still struggle to collect well-structured and high-quality datasets. Industrial data tend to be fragmented across different “siloed” systems such as CPPS systems, business information systems (e.g., ERP (Enterprise Resource Planning) system), asset maintenance systems (e.g., Computerized Maintenance Management System (CMMS)), historian databases, automation systems (e.g., Supervisory Control and Data Acquisition (SCADA) systems) and more. These systems comprise structured (e.g., sensor data), semi-structured (e.g., operations records) and unstructured data (e.g., product images), while having different semantics, formats, and representations. Likewise, some of the data have high velocity (e.g., sensor data), while others are simply high-volume data at rest (e.g., transactional data). The above-listed factors make their integration and utilization in AI use cases very challenging. To alleviate these interoperability and integration challenges, several interoperability and BigData management technologies have been developed, including technologies based on industrial standards for data interoperability. Nevertheless, even with interoperability solutions at hand, industrial organizations must also confront one more challenge, namely the unreliability of industrial data.

Data reliability is one of the most important issues in industrial environments, especially when it comes to developing and deploying AI systems. Without reliable data, it is almost impossible to develop effective and high-performance AI algorithms due to the well-known GIGO (Garbage In Garbage Out) phenomenon [7]. Furthermore, the use of unreliable data for training AI algorithms is likely to lead to biased and unreliable systems. There are many reasons why data reliability is very challenging in industrial environments. Specifically, industrial data collection is an inherently unreliable process because of:

- Environmental influences like high or low temperatures, humidity, moisture, and air pressure factors.
- Background noise such as noise pollution or interference (e.g., alarms, extraneous speech) and electrical noise from devices like motors, cooling devices, air conditioning, and power supplies.
- Faulty or inaccurate sensors i.e., sensing systems with poor precision.
- Dying battery of a system that compromises its ability to operate properly and provide reliable measurements.
- Compromised or attacked devices that produce biased or fake data due to adversarial attacks (e.g., data modification, false information injection).
- Compromised AI or BigData analytics algorithms, such as algorithms under poisoning or evasion attacks [8].

To alleviate data unreliability, industrial enterprises take various measures, including data cleansing of databases, quality control on the various data sources, as well as reconciliation of conflicting information towards a single version of the truth. In the area of data security and resilience, it is important to ensure that data infrastructures are cyber-resilient and cannot be tampered. In this direction, the use of distributed ledger technologies (most notably blockchains) is suggested in several research works. Blockchain technology is widely known as the underlying technology of blockbuster crypto-currencies (e.g., Bitcoin, Ethereum) and other crypto-assets. Nevertheless, it is increasingly used in other applications beyond digital finance, including industrial applications in sectors like manufacturing, energy, agriculture, and supply chain management [9].

Blockchain infrastructures provide the means for decentralized data management, including ways for decentralized data operations and transactions, such as CRUD (CReate Update Delete) operations. They come with many compelling properties for reliable data operations including:

- **Decentralized operation – No single point of failure:** Blockchain infrastructures operate in a highly distributed fashion and do not rely on a trusted third party for the validation of data transactions. This architectural property makes them much more difficult to be compromised, as they have no single point of failure. It also ensures their around-the-clock availability.
- **Tamper-resilience:** Distributed ledgers feature anti-tampering properties. Data written in a distributed ledger requires a next-to-impossible investment in resources to be changed. Depending on the algorithm employed to achieve consensus within a blockchain network, there are no good incentives to attempt to tamper the state of the blockchain. This is a foundation for data reliability, as blockchain data cannot be changed by adversarial parties.
- **Data transparency and auditability:** Transactions persisting (meta)data on a blockchain are transparent and accessible to all members (peers) of a blockchain network. As such they are auditable and open to scrutiny of other participants to the distributed ledger infrastructure.
- **Security:** Blockchain infrastructures offer very secure integrity protection mechanisms, including data hashing and cryptographical linking among the various blocks. This boosts their tamper-proof nature and minimizes security risks. Likewise, it is not possible to hack a blockchain by attacking few of its nodes. Blockchains support consensus mechanisms (e.g., majority voting), which require an absolute majority of nodes to agree on changes to the blockchain contents. In this way, they are resilient against cyber-attacks that could compromise one or more nodes.

These properties make blockchain infrastructures appropriate for the implementation of manufacturing use cases that involve decentralized control, such as decentralized automation and distributed data analytics processes [10, 11]. Furthermore, blockchain technologies are ideal for implementing supply chain management use cases, given the decentralized nature of industrial value chains and the fact that they cannot always operate based on trusted third parties [12]. Also, blockchains are very appealing when it comes to implementing data provenance functionalities, which are among the main pillars of data reliability in industrial environments. In this context, the present chapter provides the following contributions:

- It reviews the use of blockchain technology in manufacturing, through providing a taxonomy of the most prominent manufacturing-oriented use cases. In doing so, our review complements other recent reviews (e.g., [9, 13, 14]).

- It provides a detailed presentation of blockchain systems for data provenance and traceability in industrial environments. As already outlined, data traceability functionalities are key to ensuring data reliability.
- It introduces a novel blockchain-based approach for AI algorithms and analytics traceability in manufacturing environments. The presented approach goes beyond the traceability of industrial data entities to the provenance of AI algorithms meta, as well as of their outcomes (e.g., analytics and classification outcomes of machine learning techniques). As such it is well suited for supporting trusted AI use cases in manufacturing, which is the overall theme of the book and of subsequent chapters.

To provide the above-listed contributions, the remaining of the chapter is structured as follows: Section 1.2 following this introduction provides an overview of the main use cases in manufacturing. The section ends-up illustrating why data provenance is one of the most important use cases for manufacturing deployments. Section 1.3 reviews the main data provenance and traceability solutions that have been proposed and/or implemented in the research literature. It focused on manufacturing applications, yet some examples from other industrial sectors are also given. Section 1.4 introduces the AI algorithms metadata and AI analytics traceability system. It presents the structure of the blockchain that is used for the implementation of the system, along with the main data and metadata tracked. Section 1.5 is the final section of the chapter. It draws main conclusions and illustrates the connection of the present chapter to other parts of the book.

1.2 Blockchain Applications in Manufacturing

1.2.1 Overview

Blockchain technology is one of the digital enablers of smart manufacturing. It is suitable for applications that require reliable, transparent and secure traceability of data, including: (i) Traceability of industrial data as part of data provenance use cases; (ii) Tracking distributed manufacturing resources towards streamlining production processes that involve multiple actors in the manufacturing chain; (iii) Secure data sharing in the scope of processes that involve exchange of digital models such as Additive Manufacturing (AM) use cases; (iv) Tracking and tracing industrial assets in support of processes like maintenance, repairs, and lifecycle assessment; (v) Ensuring transparency and auditability of manufacturing chains, including coordination of supply chain management and logistic processes; (vi) Increasing the cybersecurity of digital manufacturing infrastructures through eliminating single points of failure and enabling faster updates of IoT devices (e.g., firmware updates, patching); (vii) Development of smart diagnostics and self-service applications for

Table 1.1. Value propositions of blockchain use cases in manufacturing.

Use Cases Type	Value Propositions
Decentralized Automation	• Reduced Latency for edge computing operations close to the field (automation, analytics) • Trusted data sharing between cloud providers, manufacturers and devices
Secure Information Sharing	• Auditability and transparency of information. • Information consistency based on consensus mechanisms. • Accelerated access to information from the best peer.
Additive Manufacturing	• Secure storage of digital assets. • Trusted and transparent information sharing for IP protection.
Equipment Authentication	• Signing and sealing the interactions of equipment with data. • Ensuring transparency and auditability in the use of assets.
Cybersecurity	• No single points of failure – more difficult to hack. • Automated update of patches and firmware updates through smart contracts.
Conformity to SLAs, Standards and Regulation	• Automated lifecycle stages: discovery and negotiation, deployment, monitoring, billing/penalty, and termination. • Augmented clarity by univocally defining the rules and by recording interactions between physical and non-physical parties in a definitive manner. • Trust in environments where parties do not need to cultivate and maintain relationships of trust among them.

machines, where the machines themselves will be able to monitor their state, diagnose problems, and autonomously place service, consumables replenishment, or part replacement requests to the machine maintenance vendors; (viii) Monitoring conformity to Service Level Agreements (SLAs), Standards and Regulation by translating their mandates into self-imposed smart contracts; (ix) Enabling equipment management, and more specifically identity authentication and authorization; and (x) Providing a framework to largely automate the subprocesses that compose a maintenance equipment leasing procedure, such as two-part negotiation, payment accomplishment and insurance agreement. An overview of relevant use cases and their benefits are presented in Table 1.1.

Following paragraphs present the main blockchain use cases in manufacturing environments.

1.2.2 Decentralized Manufacturing Automation: Intelligence Beyond Cloud-based Manufacturing

Cloud computing is nowadays considered an integral element of most Industry 4.0 deployments. Many manufacturing use cases integrate and analyse data on the cloud to enable applications like asset management and digital twins. This cloud-based approach to digital automation has significant advantages stemming from the scalability, capacity and quality of service offered by cloud infrastructures. Nevertheless, it comes also with disadvantages such as: (i) The need to transmit data over a wide area network that results in high latency and is not appropriate for low latency use cases that involve automation operations close to the field; and (ii) the requirement for continuous internet connectivity, which cannot be taken for granted in industrial environments.

Blockchain technology can alleviate the challenges of cloud-based manufacturing through enabling decentralized automation approaches that reduce latency and boost smartness. As a prominent example, in [15] and [10] the authors leverage blockchain technology to implement an edge computing automation paradigm, including distributed automated and distributed data analytics functionalities. As another example, in [16] a blockchain has been used to speed up the flow of production operations based on the coordination of information flows across manufacturing plans and warehouses by means of CPS systems and IoT devices. It falls in the scope of a broader class of systems that aim at deviating from cloud-based manufacturing towards supporting real-time interactions with CPPS systems. Interconnectivity between devices is considered important in this direction [17].

When compared to cloud-based automation systems, blockchains provide also increased trust between interacting actors, including users, CPPS systems and shopfloor services. Specifically, they facilitate trusted data sharing on the shop floor level. This is outlined in [18], which proposes two complementary blockchain-based infrastructures for data sharing: (i) A public blockchain network that is destined to facilitate trusted interactions between cloud providers and manufacturers and (ii) A private blockchain network that boosts trusted data sharing at the shop level, leveraging machine-level connections for data collection.

1.2.3 Secure Information Sharing

Blockchain infrastructures can be used to establish a trusted decentralized environment for sharing data in a secure way. This fosters the implementation of secure information sharing use cases. For instance, a blockchain based system that boosts information sharing for Injection Mould Redesign (IMR) is described in [19]. The system emphasizes trusted information sharing between blockchain participants, while at the same time optimizing the efficiency of the sharing processes through selecting the most appropriate (i.e., faster) peer for accessing the shared knowledge.

In another trusted information sharing use case, a blockchain infrastructure enables the implementation and execution of smart contracts for the sharing of critical information [20]. The blockchain network comprises peers deployed in manufacturing machines, system-on-chip platforms, and computing nodes. These peers enable a consortium of disparate organizations to communicate through a decentralized network. Trust is boosted by the application of data provenance mechanisms based on a proper audit trail.

1.2.4 Additive Manufacturing

The implementation of blockchain infrastructure for secure and trusted data exchange is particularly useful for AM applications [21]. The latter leverage digital models of the products and are constantly gaining momentum in the scope of the Industry 4.0 revolution. This is because they improve product design, boost shorter time-to-market, and increase manufacturing agility. One of the biggest challenges of AM is the secure exchange of data across the stakeholders involved in the production process, also given the fact that some of the exchanged data comprise Intellectual Property (IP) (e.g., digital models of a product) and other valuable assets. Blockchain technology facilitates the secure storage and exchange of digital assets, which is the reason why there many blockchain use cases in AM.

In a recent research paper [22], blockchain technology has been used for IP rights management in the context of an AM network. It has been also exploited for monitoring printed parts through their lifecycle, while tracking process improvements. The solution is studied in a broader supply chain context, where the benefits of blockchain (i.e., security data sharing, enhanced visibility, and auditability) are highlighted and acknowledged. Another blockchain application for AM is presented in [23], which emphasizes in the metal additive manufacturing process for components of the aircraft industry. The application is essentially a digital twin that supports the secure end-to-end tracing of data generated during AM processes.

1.2.5 Equipment Identity Management

Distributed ledger technologies are well-suited to provide an effective mechanism that enables equipment management, and more specifically identity authentication and authorization. Quality control requires strict supervision over which equipment has clearance to modify which subsets of data collections. What is more, the logs assembled during the procedure ought to be immutable. By assigning a digital identifier to each piece of equipment (e.g., a sensor in an IoT infrastructure), allowing it to univocally “sign” its interactions with data, transparency and, therefore, an uncontested single source of truth are formulated step by step [24, 25].

In practice, various blockchains and other DLTs provide the possibility of creating unique accounts, fitted with a pair of cryptographic keys; a public one to be universally authenticated and a private one to "sign" transactions. In such a configuration, any interaction with data is signed with the equipment's private key and can be verified by anyone who has access to the latter's public key. This verification proves that the equipment had access to the private key, and therefore is likely to be the one associated with the public key. This also ensures that the digital signature has not been tampered with, as it is mathematically bound to the key it originally was made with. From their part, smart contracts can be employed for both handling authorization requests and translating authorization policies into machine-readable self-executing code.

Overall, the use of DLT for equipment identify management, provides the following benefits:

- Security risks related to password authentication are mitigated. For example, there is no possibility for the third party to use a simple/frequent password or to share it unintentionally.
- Authentication of mobile devices, such as phones, tablets, and Augmented Reality (AR) glasses, is less prone to risk. No cookies or other retrievable objects remain on the device.
- Storage and logs are immutable per DLT specifications.
- Weak authentication protocols and human negligence do not pose a threat.
- Security regulations are more severe and privately manageable as opposed to cloud repositories.
- There is no centralized data honeypot for hackers to target.
- There is no need for action if the external user for some reason needs to be un-certified in the future.

1.2.6 Cybersecurity

Many of the previous listed use cases come with security-related value propositions. The latter emerge indirectly e.g., as part of securing data sharing processes. However, blockchain technologies can be used for strengthening the cybersecurity of digital manufacturing infrastructures [14]. For instance, blockchain technology boosts application decentralization, which eliminates single points of failure and boosts the distributional of computational loads across various servers. Likewise, distributed ledger technologies enable decentralized ways for automating the process of updating or patching IIoT devices based on smart contracts [26]. They are also used to provide decentralized trust and accountability without relying on trusted parties [27]. Furthermore, using blockchain technologies data from IIoT devices can be anonymized and remain private within edge nodes i.e., hardly accessible to non-authorized users.

1.2.7 Monitoring Conformity to Service Level Agreements (SLAs), Standards and Regulation

The advantages of translating Service Level Agreements to self-imposed smart contracts are noteworthy [28]. First and foremost, this process automates their lifecycle stages: discovery and negotiation, deployment, monitoring, billing/penalty and termination. Furthermore, it introduces clarity, since all rules are univocally defined, and transparency, since all interactions between physical and non-physical parties are recorded in a definitive manner. Lastly, DLT technologies are suitable for environments where parties do not need to cultivate and maintain relationships of trust among them [29].

In a real-world application, a "master" smart contract can be designed to enforce legal standards and agreements of any kind. By cross-examining the data uploaded by different stakeholders all parties can verify to what extent the process meets the predefined regulatory conditions. Once all the requirements are met, the regulatory approval may be automatically granted through a smart contract with no further need for on-site inspections or in-person verification.

1.3 Blockchain-based Data Provenance and Traceability

1.3.1 Overview

Data provenance and traceability is one of the most prominent blockchain use cases in industry. It is commonly proposed in cases where several of the following issues hold:

- **Resilience Concerns:** Centralized data provenance databases are more susceptible to hacking and sometimes do not withstand failures. This motivates the use of more decentralized infrastructures.
- **Multiple Writers in different Trust Domains:** There are multiple writers in the data provenance and traceability database, which may raise trust issues. This is particularly relevant in cases where writers belong in different administrative domains and trust domains.
- **Lack of clear rules to Control Data input:** Blockchain infrastructures are suitable for implementing data input rules (including validation rules) by means of consensus mechanisms and smart contracts [30].

1.3.2 Data Provenance Systems

One of the first and most prominent blockchain-based data provenance infrastructures is Provchain [31]. It enables auditing of data operations over cloud storage in real-time, while supporting access control and intrusion detection. The infrastructure leverages the tamper-proof properties of blockchain technology by maintaining a decentralized time-stamp log of user operations along with a blockchain receipt. Moreover, it supports privacy preservation features by preventing a direct correlation between users and provenance records, by means of a hashed user ID. Finally, Provchain validates provenance data entries by confirming every block using a blockchain receipt.

There are also solutions that provide traceability and provenance at the level of entire data entities. For instance, the ProductChain blockchain [32] aims at keeping track of processes in the food supply chain. In this direction, a permissioned blockchain is employed along with a transaction vocabulary for the target domain. The system provides interfaces that enable consumers and other stakeholders to access food product provenance information, without disclosing information about trade flows. ProductChain provides very good performance (i.e., query response times) and is therefore suitable for a broader class of supply chain management applications. Beyond product data provenance and traceability in the food chain, there are also blockchain systems for manufacturing chains (e.g., composite materials traceability) [33] and other agricultural products [34]. The latter are recorded in terms of their identity, species name, planting-time, company-name, greenhouse number, and geographical location. Likewise, provenance records about agricultural processes include information about identity, date & time, person, digital-signature, location, operation type, and company.

As another example, the SmartProvenance system leverages smart contracts and consensus mechanism [35] to provide reliable data provenance assuming that the majority of blockchain participants operate properly. The system supports privacy preservation using public key encryption and digital signatures. Blockchain data provenance infrastructures are commonly combined with cloud infrastructures to provide traceability of metadata such as industrial processes configuration information [36].

1.3.3 Gaps for Trusted AI

The above-listed provenance systems provide a sound basis for the implementation of data traceability platforms for reliable BigData analytics in industrial applications. Nevertheless, they are mostly focused on tracing data entities like products and assets without provisions for the provenance of machine learning and deep learning algorithms. The latter is important for the implementation of consistent, reliable, and trusted AI analytics operations in manufacturing environments. AI/ML provenance can help detecting and mitigating cyber-attacks against AI systems such as poisoning. It can also boost the implementation of configurable cyber-defence strategies that are grounded on auditing the trustworthiness of AI/ML algorithms training data. The traceability of AI algorithms by means of blockchain infrastructures and smart contracts has a dual flavour:

- **Provenance of AI algorithms metadata and configuration:** Blockchain infrastructures can be used to boost the integrity of AI algorithms and models configurations such as the weights of the neurons of a deep neural network or the parameters of linear regression algorithms. In a trusted AI environment, reliable algorithms that have not been hacked must be used.
- **Provenance of AI analytics outcomes:** Many manufacturing decisions are based on the outcomes of analytics operations. To this end, malicious parties may attempt to compromise the integrity and correctness of analytics outcomes. Blockchain infrastructures can boost the correctness and consistency of analytics outcomes to ensure that manufacturing operations leverage the actual outcomes of trusted AI algorithms i.e., that the outcomes are trusted as well.

1.4 Blockchain Data Provenance for Trusted AI in Manufacturing

1.4.1 Overview and Scope

The EU funded STAR project[1] researches, develops and validates technologies that boost trusted AI in production lines. The project studies technologies that cover many different AI systems in the manufacturing sector, including machine learning and deep learning algorithms, as well as human robot collaboration. It also deals with the safety of these systems such as the safe movement of autonomous mobile robots during their operation in a production plant. The scientific and the technological development areas of the STAR project are presented in other chapters of the book.

Data reliability is an integral element of trusted AI systems in manufacturing, as they are needed for training and operating trusted AI systems. To ensure industrial data reliability, STAR develops a blockchain-based data provenance infrastructure. The latter is destined to leverage the benefits of distributed ledger technologies that were presented in the previous section. The STAR blockchain is not intended to substitute conventional battle-hardened databases. This is because blockchain infrastructures are not best suited for managing large volumes of data

[1] H2020 STAR (Safe and Trusted Human Centric Artificial Intelligence in Future Manufacturing Lines) Project, Grant Agreement Number 956573, https://star-ai.eu/

and transactions. Customization and frequent changes are among blockchain's top enemies. Furthermore, blockchain cannot compete with conventional databases and datastores in performance (e.g., in terms of latency) and responsiveness. Finally, blockchain infrastructures are not suitable for use cases that require reversibility (e.g., rollback operations). Considering the above-listed limitations, the STAR blockchain is implemented as a data infrastructure that complements other data management infrastructures (e.g., databases, datastore) in persisting and manage provenance data (i.e., "meta-data") about data entities like AI algorithms and their analytics outcomes.

Overall, the STAR data management infrastructures exploit the best of both worlds:

(i) State of the art BigData management infrastructures are used to persist large volumes of raw transaction data and to support data operations over them;

(ii) A blockchain infrastructure is used to persist metadata about industrial data towards offering data provenance and traceability functionalities for industrial data entities, including AI algorithms and their outcomes. These metadata boost the cyber-security of the systems and enable the implementation of security risks mitigation strategies.

1.4.2 Performance Considerations and Blockchain Selection

Even though the STAR blockchain is not destined to store and manage large volumes of raw data transactions, it must feature a decent performance for the provenance tasks at hand. Specifically, a blockchain for industrial metadata provenance should provide support for several hundreds of transactions per second, in order to persist the metadata and the outcomes of the AI models that are deployed and executed in a factory.

Public blockchain infrastructures are usually criticized about their poor performance. For instance, the bitcoin protocol is one of the slowest blockchain protocols as new blocks of transactions are validated every ten (10) minutes on average. Ethereum is much faster (i.e., approx. 15 seconds per block are required), yet far from providing decent performance for industrial use cases. This slow performance of public blockchains is due to their complex mining algorithms that safeguard their high security and provide the means for generating new cryptocurrencies. Likewise, this slow performance comes with poor energy efficiency, as mining algorithms are energy intensive.

While public blockchains are not suitable for industrial applications, there is another class of blockchain networks that provides performance suitable for manufacturing use cases. Specifically, this is the case with private permissioned blockchain networks where anyone can participate in the distributed ledger infrastructures as soon as it has proper permission from the governing entity of the blockchain. In this blockchain type, the governing entity of the blockchain defines the operations that each one of the participants can perform on the blockchain in terms of creation and execution of transactions and smart contracts. Permissioned blockchain networks consist of a controlled and more limited number of participants. As such they need not operate based on complex mining protocols and Proof-of-Work (PoW) but can rather dispose with lightweight consensus protocols such as Proof-of-Stake (PoS) [Bashir18]. Permissioned blockchains remain slower than ordinary databases and do not scale to 100s of thousands of transactions per second, yet they can achieve performance of few thousands of transactions per seconds. Considering the requirements of the STAR data provenance system, we opted for a private permissioned blockchain. Moreover, we selected the HyperLedger Fabric infrastructure for the implementation of the STAR blockchain [37].

1.4.3 Architecture of Data Provance and Cybersecurity Sub-System

A part of the STAR architecture for secure and trusted AI systems is depicted in Figure 1.1. Specifically, the figure illustrates the sub-system that deals with the reliability of industrial data and the security of AI algorithms in industrial environments. The sub-system sits between: (i) The digital manufacturing platforms and the CPPS systems of an Industry 4.0 shopfloor and (ii) The security teams of the factory, such as Factory IT and cybersecurity experts, as well as CERTs (Computer Emergency Response Teams) and CSIRTs (Computer Security Incident Response Teams). The architecture specifies the following modules:

- **Runtime Monitoring System (RMS) – Data sources connectors and probes:** The RMS provides the means for collecting information from the CPPS systems and the digital manufacturing platforms of the shopfloor. It comprises several configurable data sources connectors and probes, which perform the data acquisition. Connectors and probes are also in charge of capturing the metadata associated with each data source and data capture (e.g., the identifier, the type, and the data formats of the source).
- **Data provenance and traceability applications:** These are decentralized applications (i.e., smart contracts) that run over a permissioned blockchain. They write metadata (i.e., industrial data and AI algorithms configurations) in the various peers of the blockchain. Moreover, they access data configurations that are written in the blockchain, including information about data volumes, statistical data properties, data locations etc. of various data sources. Such decentralized applications are used by the cyber-defence strategies of the sub-systems to identify possible security risks associated with AI algorithms.

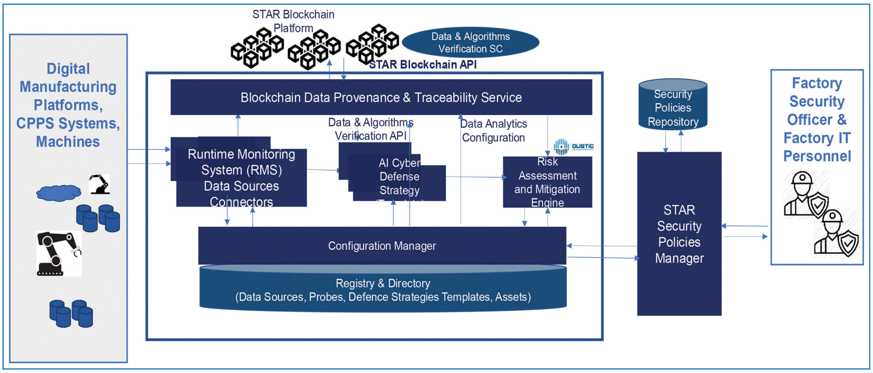

STAR Blockchain Platform
Data & Algorithms Verification SC
STAR Blockchain API
Digital Manufacturing Platforms, CPPS Systems, Machines
Blockchain Data Provenance & Traceability Service
Data & Algorithms Verification API
Data Analytics Configuration
OLISTIC
Runtime Monitoring System (RMS) Data Sources Connectors
AI Cyber Defense Strategy
Risk Assessment and Mitigation Engine
Configuration Manager
Registry & Directory
(Data Sources, Probes, Defence Strategies Templates, Assets)
Security Policies Repository
STAR Security Policies Manager
Factory Security Officer & Factory IT Personnel

Figure 1.1. Part of the architecture of the STAR system for secure and trusted AI in manufacturing.

- **AI Cyber-defence strategies:** These are security mechanisms that detect security risk and attacks against AI systems, such as poisoning and evasion attacks. They are structured in the form of templates that can be contextualized to different manufacturing environments. In detecting the various attacks, they leverage information about data configurations that are persisted in the blockchain. To this end, they access smart contract functionalities through an appropriate façade and a related API (Application Programming Interface). For instance, a cyber-defence strategy may use information about the statistical properties of the industrial datasets that are used to train an algorithm, to identify a potential poisoning attack. In this case, the blockchain will provide the actual "sealed" statistical properties, which will differ from the poisoned training data.
- **Risk assessment and mitigation engine:** This module accesses the importance of detected risks and proposes actions for their mitigation. It comprises a Security Knowledge Base (SKB) i.e., a repository of known vulnerabilities and attacks. The engine consults the SKB towards providing fast revolution of known attack patterns, as well fast identification of related mitigation actions.
- **Registry of probes, algorithms, templates, and other assets:** To support the configurable and dynamic operation of the system, the various components are registered in a proper catalogue (i.e., registry). The registry provides real-time information about the probes, algorithms and templates that are available, along with information about their status (e.g., active, inactive). Part of the registry is used by the RMS, as presented in following paragraphs of this section.
- **Security policies manager:** This module configures the operation of the subsystem through activating and configurating specific data sources, probes, cyber-defence strategies and AI models. These configurations are provided

in the form of a security policy, which is activated and enforced based on interactions with the above listed modules. Security experts and teams are responsible for specifying and deploying proper security policies.

1.4.4 Blockchain Network Implementation and Deployment

Figure 1.2 illustrates how the Blockchain Data Provenance and Traceability service interacts with other non-Blockchain modules of the STAR platform.

It exhibits a rather complex architecture, the assemblage of which requires the use of several interconnected machines each hosting some of its components, thus formulating a private permissioned Blockchain network. An Organization participating in the network in this context is a non-Blockchain module of the STAR architecture, such as the RMS or the Configuration Manager, that gains benefit from recording information on the Blockchain. Everything that interacts with the Blockchain network acquires their organizational identity from their digital certificate and their Membership Service Provider (MSP) definition.

Communication of service owners with the Blockchain Network, takes place indirectly, but via a multi-level Backend application that exposes several APIs to client applications. Another choice would be to transfer those functionalities directly to the platform's various service components, which conforms more

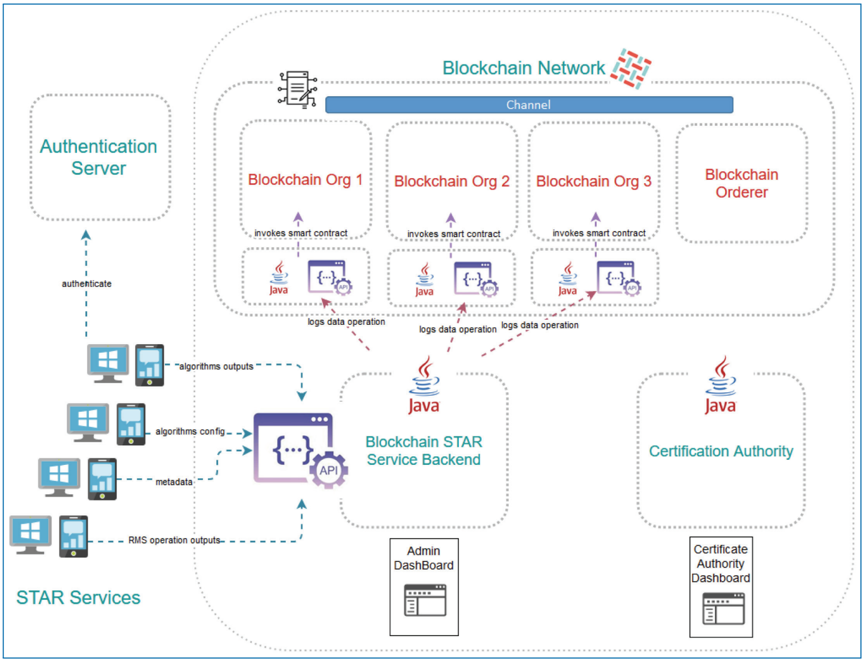


Figure 1.2. STAR data provenance and traceability service.

naturally with the Blockchain decentralization paradigm, but this would have required their developers to have extensive expertise in decentralized applications development. A final proposal would be to make Smart Contracts handle everything a back-end process does, including the job of the APIs. However, assigning only specific tasks to each distinct component has been judged to be way cleaner and manageable.

The various APIs serve different users: one exists for the Authority tasked with producing the certificates that will allow participation on the network. Another serves administrative and monitoring tasks. The most important API serves the parties recording and retrieving data. To conclude, users are authenticated against an identity management server (for instance Keycloak[2]), which entitles them to access the permissioned Blockchain network.

Figure 1.3 provides an anatomy of the Permissioned Blockchain network implementation of the project, shedding light into the various technologies used to materialize its components. The building blocks and operational processes of the Blockchain are directly dictated by the Hyperledger Fabric architectural paradigm. Specifically, the architecture exhibits a two-levels structure: (i) A first level that comprises different administrative entities (i.e., modules synthesizing the STAR platform) and (ii) A second level that comprises various peer nodes (i.e., sub-components of said modules) within each service. In-line with the Fabric's architecture, the various peers can interact and exchange data through one or more Channels. Only the peers that participate in a Channel can communicate through it and share joint ownership of the information stored on the Blockchain. This provides flexibility in clustering the peers in different groups that engage in various groups of disjoint transactions. Every node can participate in several Channels i.e., it can communicate with different groups of peer nodes. One or more Smart Contracts, which in the STAR context are describing traceability information and algorithm configurations, are deployed across a Channel.

Each (peer) node maintains a ledger of the transactions where it is involved. To this end, each peer maintains a database such as Apache CouchDB.[3] Every time the global state commonly maintained via the Blockchain changes (e.g., new metadata about a data source become available) a new transaction is initiated, through an interaction with a Smart Contract. The latter is responsible for consistently changing the status of said state to reflect the inclusion of the new information. In the Fabric infrastructure, some nodes can propose and endorse transactions, while others are only able to propose them. Nodes that can endorse transactions ought to

[2] Keycloak Open Source Identity and Access Management System, https://www.keycloak.org/

[3] Apache CouchDB, https://couchdb.apache.org/

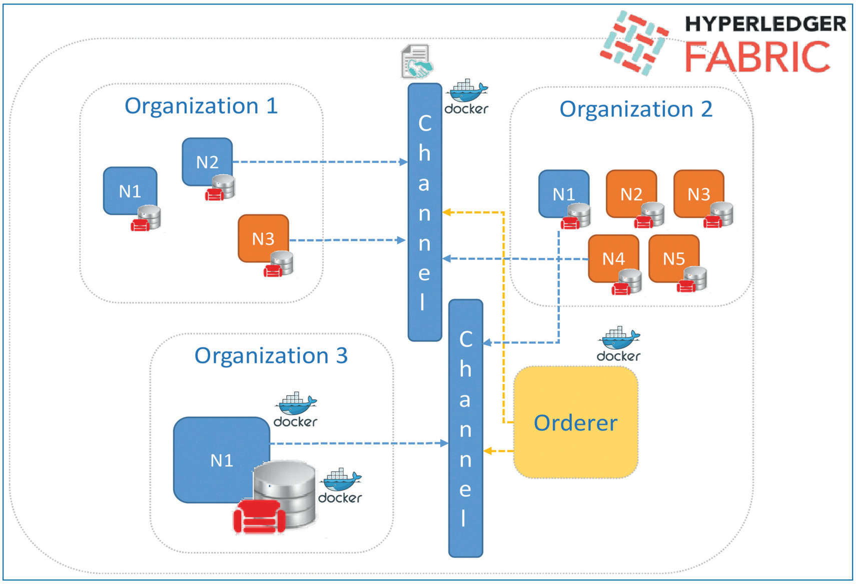


Figure 1.3. STAR permissioned blockchain network example.

reach an agreement on the current global state of the data by employing a consensus algorithm i.e., Raft[4] in our case. This increases the flexibility of the permissions and functionalities that can be granted to the different nodes, which represent a data source within digital manufacturing platforms and CPPS systems.

As specified in the Fabric paradigm, special nodes (called "Orderers") validate the various requests to update the state against the existing configuration, generate new configuration transactions, and package them into blocks that are relayed to all peers on the Channel. The peers then process the configuration transactions in order to verify that the modifications approved by the Orderers do indeed satisfy the policies defined in the Channel.

1.4.5 Data Modelling and Persistence

To implement industrial data provenance, there is a need for using a proper data model of the industrial metadata that will be stored in the Blockchain. In this direction, STAR extends background digital models of the authors [38], which comprise the following main metadata:

- **Data Source Definition (DSD):** Defines the properties of a data source on the shop floor, such as a data stream from a sensor, a CPPS, or an automation device.

[4] Raft Consensus Algorithm, https://raft.github.io/

- **Data Interface Specification (DI):** The DI is associated with a data source and provides the information need to connect to it and access its data, including details like network protocol, port, the network address and more.
- **Data Kind (DK):** Specifies the semantics of the data source. The DK can be used to define virtually any type of data in an open and extensible way.
- **Data Source Manifest (DSM):** Specifies a specific instance of a data source in-line with its DSD, DI and DK specifications. Multiple manifests (i.e. DSMs) are therefore used to represent the data sources that are available in the factory.
- **Observation:** Models the actual dataset that stems from an instance of a data source that is represented through a DSM. Hence, it references a DSM, which drives the specification of the types of the attributes of the Observation in-line with the DK that facilitates the discoverability of the data. An Observation is associated with a timestamp and keeps track of the location of the data source in case it is associated with mobile device or CPPS (e.g., mobile robot). The value type of observation is a complex object which is described with the DK entity that an Observation references. Hence, an observation can depict multiple raw measurements coming from a machine or a single value (i.e., the number of cycles/m of a rotor) or even an Analytics result (i.e., the calculated Remaining Useful Life (RUL) of a machine).
- **Edge Gateway:** Models an edge gateway of an edge computing deployment i.e., a deployment following the edge computing paradigm. In the scope of an edge computing deployment, data sources are associated with an edge gateway. This usually implies not only a logical association but a physical association as well, i.e. an edge gateway is deployed at a station and manages data sources in close physical proximity to the station.

The above entities are used to represent the data sources of a digital shopfloor in a modular, dynamic, and extensible way. This is based on the registry of data sources and their manifests, which keeps track of the various data sources that register to it. Furthermore, to facilitate the management and configuration of analytics functions and workflows over the various data sources, several analytics-related entities are also specified, including:

- **Analytics Processor Definition (APD):** Specifies a processing function to be applied on one or more data sources. Three processing functions are defined, including functions that pre-process the data of a data source (i.e. Pre-Processors), functions that store the outcomes of the processing (i.e. Store Processors) and functions that analyse the data from the data sources (i.e. Analytics Processors). These three types of processors can be combined in various configurations over the data sources in order to define different analytics workflows.

- **Analytics Processor Manifest (APM):** Represents an instance of a processor that is defined through the APD. The instance specifies the type of processors and its actual logic through linking to a programming function (e.g., Java).
- **Analytics Orchestrator Manifest (AOM):** Represents an entire analytics workflow. It defines a combination of analytics processor instances (i.e. of APMs) that implements a distributed data analytics task. The latter can span multiple edge gateways and operate over their data sources.

The digital models for industrial data provenance and data analytics follow a hierarchical structure, which defines the different relationships between the various entities. For example, an edge gateway comprises multiple data source manifests. Each one of the latter is associated with a data source definition. Likewise, Observations are associated with instances of data sources i.e. data sources manifests [Kefalakis19], [Soldatos19].

Also, the digital models presented above offer some special characteristics in order to be adaptable in various AI-based application specific scenarios such as predictive maintenance, quality management and zero defect manufacturing (ZDM). The Entities that facilitate such extensions are the Data Kind (DK), Observations and Additional Information.

Data Kind specifies the semantics of the data source data, which provides flexibility in modelling different types of data. It can be used to define virtually any type of data in an open and extensible way. It describes the type, format and data kind of the values that are produced by an AI system. Specifically, the "kind" of the data is represented with the QuantityKind attribute which is an abstract classifier that represents the concept of "kind of quantity". A QuantityKind represents the essence of a quantity without any numerical value or unit. (e.g. A sensor -sensor1- measures temperature: sensor1 has quantityKind temperature). The Data Kind is not only used to describe data sources that are used by an AI system, but also data sources that are produced from the system i.e., the outcomes of AI/ML analytics.

Observation entities model the actual data that stem from an instance of a data source (i.e., modelled through a DSM). Hence, it references a DSM, which drives the specification of the types of the attributes of the Observation in-line with the DK. An Observation is associated with a timestamp and keeps track of the location of the data source in case it is associated with a mobile (rather than a stationary) data source. Hence, it has a location attribute as well. Observation holds the measurement or result of a Data Source at the "value" entity which is of type anyType. This means that it can support any type of value (even complex structures) that are identified from the Data Kind it is referencing. Similar to Data Kind and Data

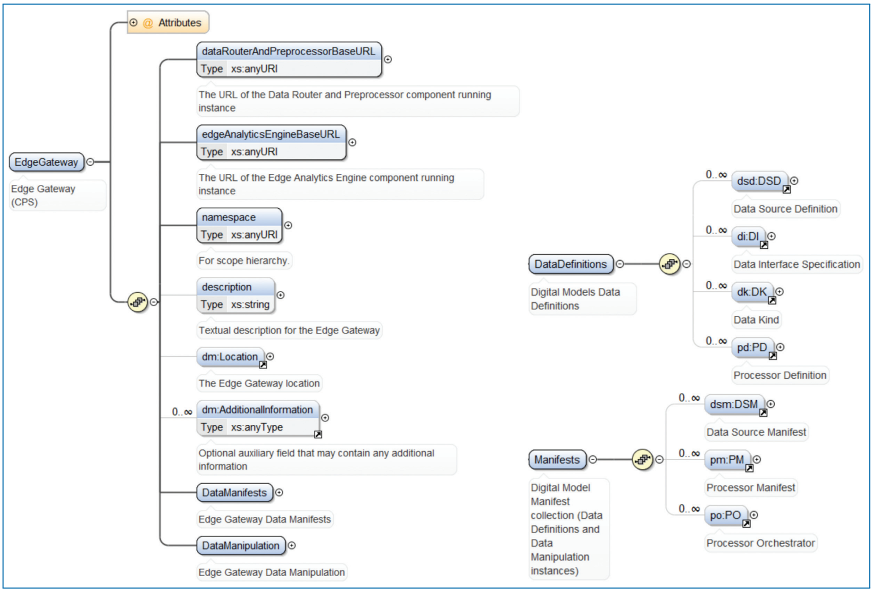

Attributes
EdgeGateway
Edge Gateway (CPS)
dataRouterAndPreprocessorBaseURL
Type xs:anyURI
The URL of the Data Router and Preprocessor component running instance
edgeAnalyticsEngineBaseURL
Type xs:anyURI
The URL of the Edge Analytics Engine component running instance
namespace
Type xs:anyURI
For scope hierarchy.
description
Type xs:string
Textual description for the Edge Gateway
dm:Location
The Edge Gateway location
0..∞
dm:AdditionalInformation
Type xs:anyType
Optional auxiliary field that may contain any additional information
DataManifests
Edge Gateway Data Manifests
DataManipulation
Edge Gateway Data Manipulation
DataDefinitions
Digital Models Data Definitions
0..∞
dsd:DSD
Data Source Definition
0..∞
di:DI
Data Interface Specification
0..∞
dk:DK
Data Kind
0..∞
pd:PD
Processor Definition
Manifests
Digital Model Manifest collection (Data Definitions and Data Manipulation instances)
0..∞
dsm:DSM
Data Source Manifest
0..∞
pm:PM
Processor Manifest
0..∞
po:PO
Processor Orchestrator

Figure 1.4. Snapshot of the digital models metadata.

Source the Observation is not only used to describe data that are captured by an AI system but also data that are produced from the system (i.e., AI analytics produce Observations).

Finally, AdditionalInformation is a generic entity which allows the extension of the existing data model with additional attributed that may be required. For instance, AdditionalInformation entities are used in the EdgeGateway and Core digital model structures to provide optional auxiliary fields that can be used for further extensions.

Figure 1.4 presents a snapshot of the different entities that are used by the different components. These are persisted in the blockchain to support provenance and traceability functionalities ate different levels and for different AI applications. For example, the Edge Gateway data entities provide discoverability of the Edge systems and enable configuration and management of the data provenance system. DSM entities provide a global and local visibility of the different industrial data sources (i.e. DSDs) used by the AI systems. This enables also discoverability of DSDs and of the Observed data that they comprise. Furthermore, the DSD can be used at configuration time to associate the data source with some AI algorithm and its outcomes. Finally, the persistence of AOM information facilitates the configuration of AI analytics functions and enables the dynamic discovery of analytics outcomes (i.e., Observations) based on the AOM they are associated with i.e., based on the AI data and algorithms that produced these observations.

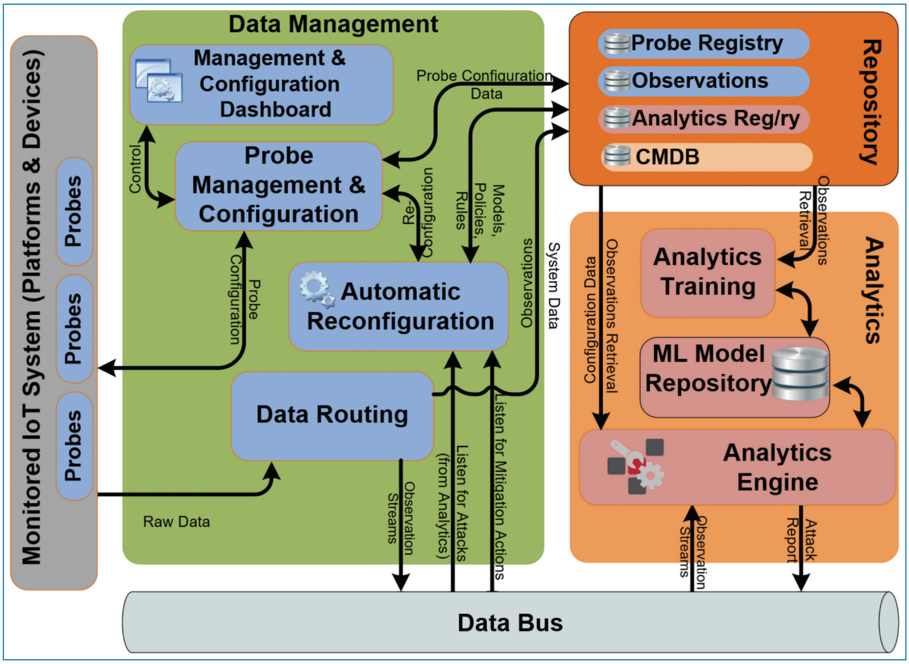


Figure 1.5. Run-time monitoring system in the trusted AI infrastructure.

1.4.6 Run-Time Monitoring System

The Runtime Monitoring System (RMS) collects industrial data in real-time and stores them for further processing and analysis by AI algorithms. RMS features lightweight monitoring probes that are responsible for the data collection and publishing to AI platforms. The RMS provides configuration and management mechanisms over the monitoring probes as well as data models and data transformation engines that will enable the discoverability and reusability of the collected data. The probe management is facilitated by an internal probe registry that maintains information about the probes (including their status), while enabling probe creation, reconfiguration, and discovery. Figure 1.5 illustrates a functional diagram of the main system components.

The main functionalities and interactions of these components are as follows:

- **Data bus:** This is a communications channel that routes real time data. Platform components may subscribe to the data bus to receive data of specific interest.
- **Deployed probes:** Probes collect data from the target IoT system or application and stream them to the IoT platform through the data routing component.

- **Probe Management and Configuration:** This module is responsible for managing and configuring the deployed probes. It can receive automatic probe configuration commands and correspondingly configures the managed probes. Manual probe configuration commands may also be received through the dashboard. The Management and Configuration dashboard provides a user interface to the Probe Management and Configuration component.
- **Probe Registry:** Maintains a record of the deployed probes. Probe deployment data, as well as state and configuration data are maintained by the registry. The registry provides probe creation, reconfiguration, and search capabilities. It facilitates the automatic deployment of probes and their dynamic discovery.
- **Automatic Reconfiguration:** This sends automatic probe re-configuration commands in-line with the implemented security policy.
- **Data Storage:** This contains historic security data that have been collected by the deployed probes. These data can be used by the Data Analytics to train itself and produce a set of security templates that will be used subsequently for identifying security issues on the target IoT system.
- **Configuration Management Database (CMDB):** This is part of the data storage. It contains information about all assets of the RMS, including their attributes and configuration parameters.

RMS is implemented based on technologies of the Elastic Stack, as outlined in Figure 1.6.

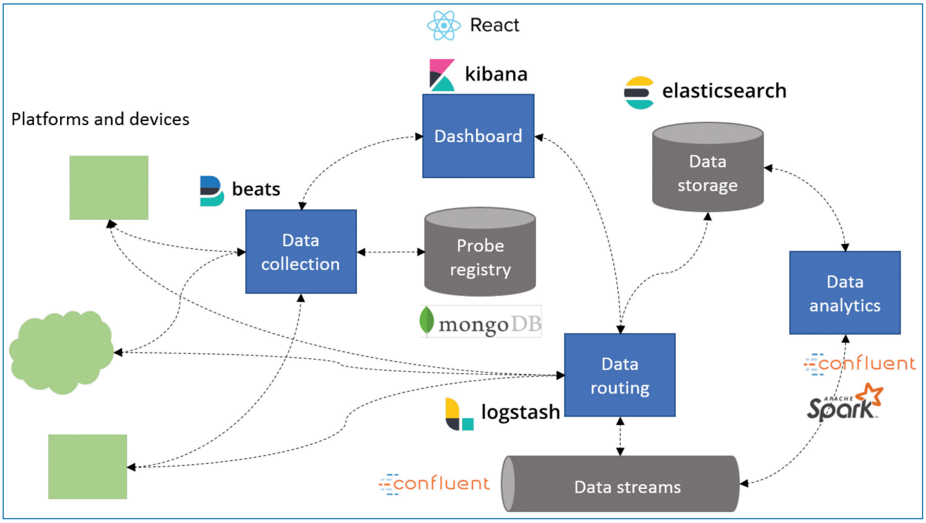


Figure 1.6. Implementation technologies of the run-time monitoring system.

1.5 Conclusion

Blockchain technologies offers advantages for data provenance and traceability in industrial environments. These advantages stem from the data security, tampered-proof and decentralized nature of distributed ledger technologies. In recent years, the research community has developed and demonstrated various blockchain-based systems for secure information sharing and data provenance in the scope of manufacturing applications. These systems provide the means for tracking industrial data entities (e.g., products, assets), as well as processes performed over them. Nevertheless, they lack support for tracking and tracing AI algorithms, models, and their analytics outcomes. This is a set-back for their use in the emerging wave of trusted AI applications. As part of this chapter, we have presented a novel blockchain infrastructure that supports data provenance for AI models and algorithms towards boosting trusted AI in industrial environments. The presented blockchain provides a foundation for data reliability. As such it blends nicely with AI systems that are presented in later chapters of the book.

The presented blockchain infrastructure is in its early implementation stages. In addition to completing its implementation and validation, we plan to benchmark its performance and scalability in real-life manufacturing environments. Moreover, we will collect feedback from manufacturing stakeholders, including practitioners with field experience. This will lead us to conclusions about the practical applicability of blockchain technology in production lines. Likewise, it will help the research community identify the next steps that could move blockchain technology from the realm of pilot experiments to practical enterprise scale deployments in industrial environments.

Acknowledgements

This work has been carried out in the H2020 STAR project, which has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 956573.

References

[1] Soldatos, J.; Gusmeroli, S.; Maló, P. and Di Orio, G. (2016). "Internet of Things Applications in Future Manufacturing", Chapter 5 in: "Digitising the Industry Internet of Things Connecting the Physical, Digital and Virtual

Worlds", editors: Vermesan, O. and Friess, P., River Publishers, pp. 153–183. ISBN 978-87-93379-81-7.

[2] Christou, I.T.; Kefalakis, N.; Zalonis, A. and Soldatos, J. (2020). "Predictive and Explainable Machine Learning for Industrial Internet of Things Applications", 16th International Conference on Distributed Computing in Sensor Systems (DCOSS), IEEE, pp. 213–218. doi: 10.1109/DCOSS49796.2020.00043. ISBN 978-1-7281-4351-4.

[3] Bai, Y.; Sun, Z.; Deng, J.; Li, L.; Long, J. and Li, C. (2017). "Manufacturing Quality Prediction Using Intelligent Learning Approaches: A Comparative Study". Sustainability, MDPI, Open Access Journal, vol. 10(1), pages 1–15, December. doi: 10.3390/su10010085.

[4] Psarommatis, F.; May, G.; Dreyfus, P.A. and Kiritsis, D. (2019). "Zero defect manufacturing: state-of-the-art review, shortcomings and future directions in research", International Journal of Production Research: 58(1), pp. 1–17. doi: 10.1080/00207543.2019.1605228.

[5] Brossard, M.; Gatto, G.; Gentile, A.; Merle, T. and Wlezien, C. (2020). "How generative design could reshape the future of product development", McKinsey & Co Article, February 2020, available at: https://www.mckinsey.com/business-functions/operations/our-insights/how-generative-design-could-reshape-the-future-of-product-development

[6] Van Roy, V. (2020). "AI Watch – National strategies on Artificial Intelligence: A European perspective in 2019", EUR 30102 EN, Publications Office of the European Union, Luxembourg. doi: 10.2760/602843. ISBN 978-92-76-16409-8. JRC119974.

[7] Geiger, R.S.; Yu, K.; Yang, Y.; Dai, M.; Qiu, J.; Tang, R. and Huang, J. (2020). "Garbage in, garbage out? Do machine learning application papers in social computing report where human-labeled training data comes from?", in Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT* '20), Association for Computing Machinery, New York, NY, USA, 325–336. doi: 10.1145/3351095.3372862.

[8] Kwon, H.; Yoon, H. and Choi, D. (2019). "Priority Adversarial Example in Evasion Attack on Multiple Deep Neural Networks", International Conference on Artificial Intelligence in Information and Communication (ICAIIC), Okinawa, Japan, 2019, pp. 399–404. doi: 10.1109/ICAIIC.2019.8669034. ISBN 978-1-5386-7822-0.

[9] Bodkhe, U.; Tanwar, S.; Parekh, K.; Khanpara, P.; Tyagi, S.; Kumar, N. and Alazab, M. (2020). "Blockchain for Industry 4.0: A Comprehensive Review", in IEEE Access, vol. 8, pp. 79764–79800. doi: 10.1109/ACCESS.2020.2988579.

[10] Isaja, M. and Soldatos, J. (2018). "Distributed Ledger Technology for Decentralization of Manufacturing Processes", proceedings of the 1st IEEE International Conference on Industrial Cyber-Physical Systems (ICPS), pp. 696–701. doi: 10.1109/ICPHYS.2018.8390792. ISBN 978-1-5386-6531-2.

[11] Soldatos, J.; Lazaro, O. and Cavadini, F. (eds.), "The Digital Shopfloor: Industrial Automation in the Industry 4.0 Era", River Publishers Series in Automation, Control and Robotics, Performance Analysis and Applications. ISBN: 978-87-70220-41-5.

[12] Abeyratne, S.A. and Monfared, R.P. (2016). "Blockchain ready manufacturing supply chain using distributed ledger", International Journal of Research in Engineering and Technology, vol. 5, pp. 1–10.

[13] Hopf, S. (2018). "Blockchain, The Emerging Platform for Manufacturing 4.0 – Major Use Cases and Implementation Challenges", The Nunatak Group, Blockchain Research Institute Paper, January 2018.

[14] ElMamy, S.B.; Mrabet, H.; Gharbi, H.; Jemai, A. and Trentesaux, D. (2020). "A Survey on the Usage of Blockchain Technology for Cyber-Threats in the Context of Industry 4.0", Sustainability 12, no. 21: 9179. doi: 10.3390/su12219179.

[15] Isaja, M.; Soldatos, J. and Gezer, V. (2017). "Combining Edge Computing and Blockchains for Flexibility and Performance in Industrial Automation", proceedings of the 11th International Conference on Mobile Ubiquitous Computing, Services and Technology (UBICOMM), pp. 159–164.

[16] Li, Z.; Barenji, A.V. and Huang, G.Q. (2018). "Toward a blockchain cloud manufacturing system as a peer to peer distributed network platform", Robotics and Computer-Integrated Manufacturing, vol. 54, pp. 133–144. doi: 10.1016/j.rcim.2018.05.011.

[17] Lee, J.; Azamfar, M. and Singh, J. (2019). "A blockchain enabled cyber physical system architecture for industry 4.0 manufacturing systems", Manufacturing Letters, vol. 20, pp. 34–39. doi: 10.1016/j.mfglet.2019.05.003.

[18] Barenji, A.V.; Li, Z. and Wang, W.M. (2018). "Blockchain Cloud Manufacturing: Shop Floor and Machine Level", Smart SysTech; European Conference on Smart Objects, Systems and Technologies, VDE, Munich, Germany, pp. 1–6. ISBN 978-3-8007-4694-1.

[19] Li, Z.; Barenji, A.V. and Wang, W. (2018). "Cloud-based manufacturing blockchain: Secure knowledge sharing for injection mould redesign", Procedia CIRP, vol. 72, pp. 961–966, 51st CIRP Conference on Manufacturing Systems. doi: 10.1016/j.procir.2018.03.004.

[20] Angrish, A.; Craver, B.; Hasan, M. and Starly, B. (2018). "A case study for blockchain in manufacturing: "FabRec": A prototype for peer-to-peer network of manufacturing nodes", Procedia Manufacturing, vol. 26, pp. 1180–1192,

46th SME North American Manufacturing Research Conference, NAMRC 46, Texas, USA. doi: 10.1016/j.promfg.2018.07.154.

[21] Diemer, M. (2019). "Blockchain – Implications and Use Cases for Additive Manufacturing", Frankfurt School Business Center, September 2019, available at: https://fsblockchain.medium.com/blockchain-implications-and-use-cases-for-additive-manufacturing-51049a644dad.

[22] Kurpjuweit, S.; Schmidt, C.G.; Klöckner, M. and Wagner S.M. (2021). "Blockchain in additive manufacturing and its impact on supply chains", Journal of Business Logistics, 2021, 42(1): pp. 46–70. doi: 10.1111/jbl.12231.

[23] Mandolla, C.; Messeni Petrucelli, A.; Percoco, G. and Urbinati, A. (2019). "Building a digital twin for additive manufacturing through the exploitation of blockchain: A case analysis of the aircraft industry", Computers in Industry, Volume 109, pp. 134–152. doi: 10.1016/j.compind.2019.04.011. ISSN 0166-3615.

[24] Omar, A.S. and Basir, O. (2018). "Identity Management in IoT Networks Using Blockchain and Smart Contracts", IEEE International Conference on Internet of Things (iThings) and IEEE Green Computing and Communications (GreenCom) and IEEE Cyber, Physical and Social Computing (CPSCom) and IEEE Smart Data (SmartData), pp. 994–1000. doi: 10.1109/Cybermatics_2018.2018.00187. ISBN 978-1-5386-7975-3.

[25] Ren, Y.; Zhu, F.; Qi, J.; Wang, J. and Sangaiah, A.K. (2019). "Identity Management and Access Control Based on Blockchain under Edge Computing for the Industrial Internet of Things", Applied Sciences. 9, no. 10: 2058. doi: 10.3390/app9102058.

[26] Fernández-Caramés, T.M. and Fraga-Lamas, P. (2019). "A Review on the Application of Blockchain to the Next Generation of Cybersecure Industry 4.0 Smart Factories", in IEEE Access, vol. 7, pp. 45201–45218, 2019. doi: 10.1109/ACCESS.2019.2908780.

[27] Locher, T.; Obermeier, S. and Pignolet, Y.A. (2018). "When can a distributed ledger replace a trusted third party?", Procedia IEEE International Conference on Internet of Things (iThings) and IEEE Green Computing and Communications (GreenCom) and IEEE Cyber, Physical and Social Computing (CPSCom) and IEEE Smart Data (SmartData), pp. 1069–1077. doi: 10.1109/Cybermatics_2018.2018.00197. ISBN: 978-1-5386-7975-3.

[28] Uriarte, R.B.; De Nicola, Rocco and Kritikos, Kyriakos (2018). "Towards distributed SLA management with smart contracts and blockchain", IEEE International Conference on Cloud Computing Technology and Science (CloudCom), vol. 1, pp. 266–271. doi: 10.1109/CloudCom2018.2018.00059.

[29] Zhang, Y.; Xu, X.; Liu, A.; Lu, Q.; Xu, L. and Tao, F. (2019). "Blockchain-Based Trust Mechanism for IoT-Based Smart Manufacturing System", in IEEE Transactions on Computational Social Systems, vol. 6, no. 6, pp. 1386–1394. doi: 10.1109/TCSS.2019.2918467.

[30] Bashir, I. (2018). "Mastering Blockchain: Distributed ledgers, decentralization and smart contracts explained", Packt Publishing Limited; 2nd Revised edition (30 March 2018), ISBN-10: 1788839048, ISBN-13: 978-1788839044.

[31] Liang, X.; Shetty, S.; Tosh, D.; Kamhoua, C.; Kwiat, K. and Njilla, L. (2017). "ProvChain: A Blockchain-based Data Provenance Architecture in Cloud Environment with Enhanced Privacy and Availability", The 17th IEEE/ACMInternational Symposium on Cluster, Cloud and Grid Computing (CCGRID), May 14–17 2017. doi: 10.1109/CCGRID.2017.8. ISBN 978-1-5090-6611-7.

[32] Malik, S.; Kanhere, S.S. and Jurdak, R. (2018). "ProductChain: Scalable Blockchain Framework to Support Provenance in Supply Chains", IEEE 17th International Symposium on Network Computing and Applications (NCA), Cambridge, MA, USA, pp. 1–10. doi: 10.1109/NCA.2018.8548322.

[33] Coronado Mondragon, A.E.; Coronado Mondragon, C.E. and Coronado, E.S. (2018). "Exploring the applicability of blockchain technology to enhance manufacturing supply chains in the composite materials industry", in IEEE International Conference on Applied System Invention (ICASI), Chiba, Japan, pp. 1300–1303. doi: 10.1109/ICASI.2018.8394531. ISBN 978-1-5386-4342-6.

[34] Hua, J.; Wang, X.; Kang, M.; Wang, H. and Wang, F.Y. (2018). "Blockchain based provenance for agricultural products: A distributed platform with duplicated and shared bookkeeping", in IEEE Intelligent Vehicles Symposium (IV), pp. 97–101, June 2018. doi: 10.1109/IVS.2018.8500647.

[35] Ramachandran, A. and Kantarcioglu, M. (2018). "SmartProvenance: A Distributed, Blockchain Based DataProvenance System", proceedings of the Eighth ACM Conference on Data and Application Security and Privacy, CODASPY '18. Association for Computing Machinery, New York, NY, USA, pp. 35–42. doi: 10.1145/3176258.3176333.

[36] Ali, S.; Wang, G.; Bhuiyan, M.Z.A. and Jiang, H. (2018). "Secure Data Provenance in Cloud-Centric Internet of Things via Blockchain Smart Contracts", IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computing, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI), Guangzhou, China, pp. 991–998. doi: 10.1109/SmartWorld.2018.00175.

[37] Androulaki, E.; Barger, A.; Bortnikov, V.; Cachin, C.; Christidis, K.; De Caro, A.; Enyeart, D.; Ferris, C.; Laventman, G.; Manevich, Y.; Muralidharan, S.; Murthy, C.; Nguyen, B.; Sethi, M.; Singh, G.; Smith, K.; Sorniotti, A.; Stathakopoulou, C.; Vukolić, M.; Cocco, S.W. and Yellick, J. (2018). "Hyperledger fabric: a distributed operating system for permissioned blockchains", in proceedings of the Thirteenth EuroSys Conference (EuroSys '18), Association for Computing Machinery, New York, NY, USA, Article 30, 1–15. doi: 10.1145/3190508.3190538.

[38] Kefalakis, N.; Roukounaki, A. and Soldatos, J. (2019). "Configurable Distributed Data Management for the Internet of the Things", Information, 2019; 10(12):360. doi: 10.3390/info10120360.

DOI: 10.1561/9781680838770.ch2

Chapter 2

Artificial Intelligence and Secure Manufacturing: Filling Gaps in Making Industrial Environments Safer

By Entso Veliou, Dimitrios Papamartzivanos, Sofia Anna Menesidou, Panagiotis Gouvas and Thanassis Giannetsos

This chapter aims to review, from the security standpoint, the artificial intelligence solutions used to empower smart manufacturing environments. Our analysis will focus on the adversarial models utilized by malevolent entities in order to cause malfunctions to AI-powered systems both during the training process, but also during the inferencing mode of the leveraged machine learning models. Such attacks can have significant impact to the operation of the manufacturing supply chain ecosystem, as they can affect not only the business continuity, but more importantly, the integrity of safety-critical operations of systems. Towards this direction, this chapter reviews the state-of-the-art in technical approaches to secure machine-learning models and pave the way towards the safe adoption of such measures in the manufacturing field. The focus is on new generation of artificial intelligence setups using at their core deep neural network structures. In addition, the chapter elaborates on attestation-based provenance mechanisms that guarantee the trustworthiness of data streams feeding AI systems. The goal is to highlight the need for robust solutions against adversarial machine learning attacks for such environments and to provide additional insights on the appropriate mitigation strategies against such intelligent aggressors.

2.1 Introduction

For many years manufacturing systems lacked information and data security, until recently that everything in the manufacturing supply chain ecosystem changed. Ethernet and IP protocol layer became the next big thing; of course, some of the driving factors for this big change were cost, need for automation and convenience. Networks became a core part of the manufacturing field and currently interconnect wider and more complex manufacturing floors. Hence, connectivity along with the increased sensing capabilities, and the desire for reduction of installation costs gave birth to an increased demand for wireless networks, multiple IoT devices, and human-robot interaction which is blooming as the new era for smart factories. The evolution of human-robot collaboration and Internet of Things have major impact on the manufacturing processes, working environment and processes, as new services can be developed by the integration of the physical and digital worlds. Moreover, this progress has an impact on the physical security of the workers and the overall safety in the smart factories, and the reason for this is because human-robot collaboration will provide to the workers a more privileged job position where the robot will handle most of the dangerous and demanding parts of the job. Smart devices and networks with improved capabilities can have significant impact on the users' well-being and on the everyday activities and procedures in a manufacturing environment with the emergence of new "systems-of-systems" (SoS).

In addition to the above, the scenery of manufacturing is rapidly changing by the penetration of artificial intelligence solutions that primarily aim to boost the productivity on the manufacturing operation process. In fact, artificial intelligence is revitalizing the smart manufacturing domain with the integration of advanced analytic methods capable of processing huge amount of data collected by the multiple IIoT devices. Based on this, predictive maintenance for minimising operation and maintenance costs, improved supply chain management, automated quality control, efficient and safe human-robot collaboration and buyer-centric manufacturing are prominent examples of added-value services that have emerged as a result of the integration of AI in the manufacturing field.

Undoubtedly, the digitisation of the manufacturing field in combination with the AI infiltration in the production processes have led to the formation of a rather complex cyber-threat landscape on smart industries. More specifically, the threats that emerge as a result of the integration of legacy ICT technologies have been widely documented in the literature [1], while several reports have documented threat taxonomies in this direction [2]. Notably, when it comes to the documentation of AI-specific threats, in other words, attacks that target specifically AI empowered systems and the leveraged AI methods, only recently the community has started to document possible attacks that can offend the operation of such systems [1, 3]. In this direction, this chapter aims to shed light on the underpinnings of the AI-fuelled smart manufacturing and in parallel to put forth adversarial techniques that can be used against such AI methods. More specifically, the focal point of this work is the in-detailed investigation of the most prominent type of attacks, namely *poisoning* and *evasion* attacks [3–6]. Poisoning attacks attempt to train the deep neural networks in ways that compromise their correct operation with the inclusion of intentionally malformed instances in the training set of AI algorithms. Evasion attacks take place at the inference stage of a deep neural network where malicious parties craft data that are incorrectly classified by deep learning systems.

In view of the above, this sets the challenge ahead: "*To which extend AI adversarial techniques can affect intelligent manufacturing systems, and what are the defensive actions that can guarantee the robustness of the AI systems towards achieving increased resilience of the production lines and business continuity?*"

Compounding this issue, Section 2.2 offers an analysis of the smart manufacturing stack by highlighting the engagement of AI solutions in the manufacturing processes. Given this analysis, Section 2.3 highlights the cyber security posture of AI-fuelled manufacturing systems by documenting impactful vulnerabilities and threats. Section 2.4 documents the importance of solutions, such as attestation that can guarantee the integrity of data flows fed into machine learning data pipelines. Section 2.5 offers a discussion and critique on the formed field's baseline before Section 2.6 elaborates on the road ahead and discuss novel solutions that can increase the residence of AI setups.

Overall, the motivation of this work is to set the scene on the need for secure AI-based systems for manufacturing environments that cannot only enable efficient decision making process but can also withstand a prolonged siege from an attacker; either targeting the integrity of the input data or the correctness of the classification model and process. Having identified the challenges and current hurdles, we also put forth a road-map of future research avenues which we need to consider if we are to fruitful benefit from the Industry 4.0 revolution.

2.2 Hardening the Smart Manufacturing Stack: Towards Inter-Trustability of System-of-Systems

Security intelligence in smart manufacturing is widely used to solve security problems, such as incident prevention, detection, and response, by applying machine-learning and other data-driven methods. The selection of intelligence sources and feeds is vast and growing, so is the choices in methods that can be applied, while the problems evolve and new ones appear. To this end, as aforementioned, there is a large body of prior work that solves security problems in specific scenarios, using specific types of data and specific algorithms [3–6]. Being specific has the drawback that it becomes hard to adjust existing solutions to new scenarios, data, or problems. Furthermore, all prior work that strives to be more general is either able work with complex relations (graph-based), or to work with time varying intelligence (time series), but never both. While there exists solutions to spatio-temporal problems in graph machine learning, they do not satisfy the conditions: 1. heterogeneity of attributed nodes, 2. time-dependence of the nodes and their attributes, 3. time-dependence of the relationships, 4. scoring of the nodes, and 5. arbitrary interactions that are not necessarily bipartite (i.e., hyperedges).

In this context, security intelligence data, or simply **intelligence**, must relate to something of relevance to security of interest, i.e., one or more specific instance of some entity types, and it must describe the entity (or entities), either through attribute(s) or by their relationship. Examples include knowledge that a device/sensor exists on the network of concern (identifies an instance, e.g., by a securely generated ID), that the device is turned on (an attribute that describes that state of the sensor), and that the device has used the Domain Name System (DNS) to resolve a domain name (Interaction between the client and domain entities).

The complete body of all security intelligence is not practically available, but parts of it can be observed. The types of intelligence we consider include also enriched observations, such as the relation between a device's ID and the hostname obtained via reverse lookup in the underlying network (programmable) infrastructure. Either way, monitoring of data is one approach to observe intelligence, which for instance network owners can use to gain insights to the traffic circulated in a smart manufacturing floor, yielding intelligence like the above. Another option is to source intelligence from others, via public or private feeds, e.g. for free or under some commercial agreement.

Whether intelligence is sourced from monitoring controlled systems, third parties, or elsewhere, the arrival of new intelligence is expected to occur at specific points in time because monitoring reveals events from observed data, or because new data from a feed arrived. To capture this, we define an **event** to be a timestamped observation of intelligence data, where an observation may for example be either a first time observation, interval since last modification or an affirmation that the previous intelligence data is still current. For instance, a data transmission from a device is an event which provides several pieces of intelligence; there is a sensor on the network that has a certain ID, it is active, and it is related to the domain name in question.

In the above, we have explained out how intelligence can be obtained from monitoring, external sources, and enrichment, but it may also be obtained from machine-learning, heuristics, manual processing and more. Common for all these processes is that they take some intelligence as input and produce some new or updated intelligence as output. This type of process we refer to as a map process, which encapsulates the knowledge of a variety of domain experts into an automated framework that enriches intelligence. In what follows, we dig into more detail behind the scenes on the types of information sources that can be considered as part of this map process; essentially, the actors that comprise this new paradigm of smart manufacturing systems that organize and integrate real-time knowledge between physical objects and the virtual computational space [8].

2.2.1 Data Source & Security Requirements of Industry 4.0: Smart Manufacturing Processes, Actors and Safety-Critical use Cases

Towards this direction, additional Cyber-Physical Systems (CPS) such as reliable indoor positioning system and activity recognition systems (e.g., motion capturing sensors), together with AI-based software solutions are among the enabling technologies that need to be leveraged. The incorporation of robotics into industrial systems has accelerated over the last decade, and there are no signals of a slowdown on the horizon. Because of regulatory and business measures, such as the German-created Industry 4.0 [7], the expanded use of robotic architectures could be an unintended result of parallel advances in a few related fields [9]. Plant systems (machines, conveyors, and so on), cognitive devices, and the cloud will both connect and share data in real time using the existing network infrastructures. Every one of the machine components, seen as units, collaborates effectively to achieve versatility and stability. Operatives must deal with problems including packet losses and ineffectiveness that may occur as a result of incompatibilities. To reduce packet loss, massive data feedback mechanisms are required [10]. Detectors play a crucial role in the application of IoT and CPS in a delivery device. A sensor is described as *a complex machine that detects light, humidity, reclamation, of some kind and sends a signal to a monitoring or controlling endpoint*. It is a good resource for converting data from the surrounding world into data in a cybernetic environment. It is proposed that self-aware and self-monitoring systems be used to capture and relay the information from the production process in actual environments [11]. When building a managed work environment with the widespread use of smart appliances, process management is often encouraged. To ensure effective communication across devices for several monitoring processes, IoT devices are further split into categories, with each class of sensors loosely deployed in a sub-area. a big factory or a long product design and development line [10]. These advancements, particularly in software engineering and automation, have allowed separate mechanisms to use smart data analysis to build process information awareness that can be used to illuminate the operational behaviour the systems and manufacturing fields.

2.2.1.1 Ideal operational requirements

The development of manufacturing advances and new processes are expected to continue in the future. Modern materials, components and objects will emerge [12]. Injection molding is an example of a modern technique that has accelerated from the innovation of modern technologies, changed the development and manufacturing of products, and unlocked the way to previously untapped areas such as biomanufacturing. Manufacturing equipment, for example, devices intended for standardized and lateral machining, as well as penetration, have been developed to manage different activities. Further type convergence will occur, such as the use of advanced products, item schedules, and production procedures, such as the identification of a chemical substance that relates to the creation of a new medication, a delivery mechanism, as well as medication production and the device. New-age robots, which are very inexpensive to build and maintain, takes smart factory automation to unpredictable levels. IoT devices and application functions make new era smart manufacturing systems more intelligent and better suited for the plant and beyond communication.

These ideal manufacturing advances in time increase manufacturing speed and productivity. Traditionally, productiveness is described to measure the degree of output as compared with a given input. Examples of inputs are individual working hours, devices hours, and materials. Productivity may be measured at unique tiers of the organizational hierarchy from an individual device to the entire organization. Productivity is outstanding from generally used overall performance goals including return-on-investment (ROI), that's a cost-primarily based frequently used at the very highest stages of the organization. A device can adjust its behaviour depending by its own knowledge with the aid of artificial intelligence, and whether it has sophisticated tracking systems, it can, for example, use cognitive computing to automate its processes and be accurate and precise. These activities and applications are susceptible to improvements in integration and may benefit from artificial neural networks. They should therefore be viewed as part of an intelligent control system. Independence exists where a device (a) may respond to feedback and act out its actions to achieve a specified goal, and (b) the unit wishes for the feedback loop to function. Advanced control technology is needed. As a result, independence must be a component of particular value. A device is said to be fully automated if it can automatically execute its own operation, although the level of automation varies from device to device [13].

2.2.1.2 Operational and performance assurance

Manufacturers usually need technological skills to monitor the range and form of technology widely available to upgrade their processes, which is posing a significant problem for industry 4.0 and smart manufacturing. Provisional application creation and evaluation are often carried out in laboratory environments, which may preclude the software from being publicized and used due to deployment challenges. This will go unnoticed by the developer. To establish that smart technological developments integrate well with traditional manufacturing processes, it is critical that the vendor and product providers collaborate to find problematic areas as well as shared solutions and best practices. To guarantee that the current framework ultimately improves efficiency, performance indicators must be identified. The use of performance enhancement standards at all stages and levels of development means that supply chains fulfil the anticipated functional criteria while also providing the appropriate guidance for quality improvement. The manufacturer's priorities must be supported by performance evidence that cascades from the highest operational level to the lowest acceptable level. It is critical that certain small indicators represent the duties given at your level while still adding to the organization's total operating measure [8].

2.2.1.3 Quality assurance

Analytical tools including simulation and statistical evaluation play a position in analysing productiveness through examination in their output reports. Advanced knowledge could also analyse comparatively existing system information, recognize correlations among differentiated system phases and inputs, and refine components that have the greatest effect on yield and productivity [12]. Replacing old fashioned manufacturing processes with Machine learning smart manufacturing processes can result in huge to slight increase in productivity and profit. Although, a reasonable question is how the quality and performance assurance are impacted from these radical changes. The quality management roadmap establishes benchmarks for enhancing quality for production processes through procurement partnerships with and within individual supplier providers. When a critical occurrence happens, it notifies human operators, allowing them to take immediate steps if possible. In case of human-robot collaboration time has taught us, that humans may be vulnerable to many types of exploits and knowledge base already exists for such type of exploitation. However, the second type of the equation is new to the manufacturing processes and various ways of exploitations can be found for a malicious individual seeking to damage the smart manufacture and attacking the machine learning algorithm behind the robot which cooperates with the human.

2.2.1.4 Control-safety and secure AI

Since the human-robot collaboration has been a core part in modern smart manufactures, as a robot we can categorize multiple IoT devices that can get involved in manufacturing processes. In that context heavy parts have to be lifted, various metallic and non-metallic components have to be machined and large plates have to be connected to one another in frequently performed tasks, big and strong devices, such as robotic manipulators, which present a severe safety threat to humans. Multiple security procedures, such as locking the machines in physical or simulated cages and holding humans at a safe range while the robots are in action, have already been introduced. However, in addition to the new conditions for modern automotive and manufacturing purposes, a new version of ISO 10218 [14], the key specification for safety specifications for robotic systems, has been created.

In the context of incorporating safety standards for autonomous or collaborative robots working with humans [12], the proposed rules for operating in a cooperative mode also include the following:

- Stopping functions (10218-1)—requirements are specified for how and when the robot should perform protective, or emergency stops when humans are in the robot's workspace [14].
- Speed and position control (10218-1)—requirements are specified for the maximum allowable speeds of robot arms and end effectors when humans are in the robot's workspace [14].
- Power and force control (10218-1)—requirements are specified for the maximum allowable power and forces applied by robot arms and end effectors when humans are in the robot's workspace [14].
- Design of collaborative operation workspaces (10218-2)—requirements are specified for the layout design of workspaces around the robot, including safeguarded spaces (where humans are separated from the robot and protected by safeguards) and collaborative spaces where humans are not separated from the robot and hence the robot shall apply the control limits [14].
- Collaborative operation modes (10218-2)—requirements are specified for the specific operating modes that must be designed into the robot's control function when collaborating with a human in the collaborative workspace, including teaching modes and autonomous modes [14].

2.2.2 Human-Robot Collaboration and IoT Devices

While the evolution of smart manufactures is radical and shifts quickly to the new era of machine learning and human-robot collaboration, the concern for physical security flourishes next to the new era. Robots and IoT devices complexity and configurations make extremely dangerous the scalability of the technologies that have been evolved within this concept. Given the clear benefits of incorporating robotics in smart manufacturing, most areas where they are being completely deployed neglect any security defense functionality by nature, making robots unreliable and vulnerable to cyber-attacks. This is one of the factors why human-robots are only preferred in testing and have not yet completely proven themselves in the market of smart manufacturing. Although it is not an easy job, many guidelines are necessary from the start to boost robot and IoT system cybersecurity [15], such as: Secure device construction development phases, encrypting robot communications, maintaining networks updated, limiting access to authorised customers, offering ways to restore a robot to a secure factory default mode, implementing cybersecurity guidance, including cybersecurity training for professional machinists and administrators, allowing consumers to provide input on potential bugs, and encouraging security assessments prior to output.

2.2.2.1 Towards trustworthy smart manufacturing processes

In smart manufacturing environments, devices can participate in the sensing process and upload their contributions to the backend (or Mobile Edge Computing (MEC) layer running) decision-making system, and raw sensor data are collected on sensor devices and processed by local analytic algorithms towards producing consumable data for requesting applications. In this context, for a specific time window with n time steps and m sensors, we consider a dataset D containing a sequence (S) for each sensor j where $Sj = [v_{1,j}, v_{2,j}, \ldots, v_{i,j}, \ldots, v_{n,j}]$.

Threat Model: The aim of adversarial agents is to mislead the smart manufacturing processes towards considering malicious measurement values as legitimate in their services. To this end, an adversary may change the input value $v_{i,j}$ in S_j to $v'_{i,j}$, where $v'_{i,j} \neq v_{i,j}$ to maximize the distortion:

$$\max\{|v_{i,j} - v'_{i,j}|\} \tag{2.1}$$

where the distortion should be lower than a maximum allowed considered by the adversarial agent.

There are two primary adversarial attack models [1, 4]: (1) pre-training (poisoning) attacks, and (2) post-training (evasion) attacks. In pre-training attacks, adversaries try to inject malicious data in an attempt to poison the training dataset and, thus, decrease the classification accuracy of the classifier. In the post-training attack scenarios, adversaries aim at misleading trained classifiers to mis-classify samples towards a malevolent intent. Let us assume $f(x_i) = y_i$ as the mapping function to calculate/map x_i to y_i . For every new sensed values x_i', f gives a new output $f(x_i') = y_i'$, and we have the following cases:

- True Positive: if x_i' is positive and f correctly outputs positive, there is no loss on the application.
- False Positive: if x_i' is negative and f outputs positive, there is a loss on the application.
- False Negative: if x_i' is positive and f outputs negative, there is a loss l on the application.
- True Negative: if x_i' is negative and f correctly outputs negative, there is no loss on the application.

In principle, a machine learning technique tries to minimize $|f(x_i') - y_i'|$ which means minimizing l and ε. On the contrary, an adversarial attacker attempts to maximize the impact of the attack by maximizing $|f(x_i') - y_i'|$.

2.3 Cybersecurity Posture of AI-Fueled Manufacturing Ecosystem

Security in smart manufacturing does not stop in the physical security of the workers. This radical change might increase safety for the workers thus it will also create a lot of information security gaps. Considering the different networking and application layers that are being involved in this big change, a lot of new vulnerabilities, attack paths, and information security gaps are being born. Considering the above threats, confidentiality and integrity must be ensured in such environments.

On the way towards such IoT-based SoS, this added richness and connectivity also poses a significant risk. The new approach of SoS will potentially leave the network vulnerable providing a huge scale of attack path to malicious users. Furthermore, in the smart manufacturing environment, this is largely underrated. Between April 2012 and January 2014, over 500,000 Computer production devices in system control ecosystems were discovered, as per Project SHINE data [16]. Since the installed smart manufacturing systems are far smaller than normal industrial equipment, it may not cause warnings to be sent to the owners of such installations because there have been relatively few attacks reported on them. However, it is worth noting that the presence of recorded attempts on such recently implemented programs does not imply a lack of vulnerabilities. It is only a matter of how long before the hacker community acquires the basic information needed to initiate successful attacks [17]. The most recent and violent assault on industrial infrastructure was the power grid attack in Ukraine in December 2015 [18]. The attackers used a combination of cybersecurity techniques such as malware, denial of service, and phishing to take the entire electricity supply infrastructure to a point where it became difficult to repair, resulting in power failures across the country. These outages caused several blackouts, affecting 225,000 clients across Ukraine. Because this incident affected the advanced manufacturing ecosystem, it is not shocking that there haven't been many accidents involving industry 4.0 systems. However, major attacks have been launched against some of the more cutting-edge smart manufacturing systems, most noticeably IoT. Relatively typical IoT nodes combine a considerably lower CPU with wireless networking network interfaces, encouraging cyber hackers to target them explicitly within their radio frequency spectrum. This contradicts the conventional security paradigm, where there is a well-defined perimeter and sensors (such as firewalls and intrusion prevention systems) are responsible for protecting the boundary. Instead, each system would have to be at least partially responsible for its own protection, a task made more difficult by the restricted processing technologies of a standard IoT node. Naturally, this is exacerbated by manufacturers failure to recognize the broad implications of inadequately securing individual devices, as well as the high-profile IoT botnet Mirai [19], which resulted in the biggest denial of service attack seen so far, is a deafening example of this disaster.

Research-wise the most promising and the one that has been given effort and developed the last couple of years is AI-based cyber defence mechanisms that are decentralized and that can more dynamically classify various attack vectors. Many efforts have been made, many algorithms have been developed and the machine learning classification models for cyber defence have gotten more sophisticated and have improved dramatically the last years. According to Sturm *et al.* (2014) [20], a void in a 3D printing component would then lead to a reduction in yield, as well as other natural physical alterations such as weight, stiffness, and attenuation coefficient. Anomaly detection can also detect unusual behaviour on a network or system (Kim *et al.* 2013) [21], as well as image (Chandola *et al.* 2009) [22], performance monitoring, and data acquisition (SCADA) (Garcia *et al.* 2011) [23], or for preventive equipment maintenance (Rabatel *et al.* 2011) [24]. It focuses on the problem of calculating the correlation that do not match expected pattern (Chandola *et al.* 2009). The concept is to identify patterns of standard practice that the algorithm has learned or indicated. Administrators will be notified if an activity deviates from the predetermined or accepted model of behaviour. When compared to existing methods, anomaly detection has the benefit of being able to detect malicious activity. That being said, the adversarial machine learning does not fall in the category where the attacker attacks the physical machine or the nodes where the AI agents are operating. In this case, the attacker tries to bypass or manipulate the classification model, which has been created, executing his real attack in a stealthy manner without being detected by the classification model. According to Kumar *et al.* (2020), It is unclear how Machine Learning vulnerabilities can be rated in terms of risk and effects. When a security specialist sees headlines of an invasion, the simple truth is usually "Is my company impacted by the attack?" and organisations today lack the intellect to search an ML area for suspected adversarial ML related vulnerabilities. In this recently adopted definition, three kinds of attacks are considered: poisoning, stealing, and evasion. The overarching aim of these models is to minimize the classification's generalization error and potentially deceive the decision-making mechanism against desirable harmful calculation metrics stated by Chen Li and Jiliang Zhang (2019) [25].

2.3.1 Poisoning Attacks

In the first scenario, the adversary will contaminate the training data. To do this, the opponent extracts and infuses an argument that reduces classification precision. This attack has the potential to totally alter the classification mechanism during training phase, allowing the attacker to interpret the system's classification in whatever way he sees fit says Vahid Behzadan and Arslan Munir (2017) [26]. The extent of the classification error rate is defined by the data used by the perpetrator to poison the preparation. The backdoor or Trojan attack, for example, is an especially sophisticated attack in this class, in which the attacker deliberately poisons the model by adding a backdoor key to ensure it performs well on normal training data and testing samples but misbehaves only when a backdoor key is used. When we are referring to model stealing, this usually can be met in confidentiality to the outer world Machine Learning models which are being implemented with an API interface that is open to the public. As an example, consider the ML as a service system: Many encourage individuals to train the models on highly sensitive data and charge others on a pay-per-query basis for use. The tension between product confidentiality and public access motivates the research of model extraction and stealing attacks. An intruder with black-box access but no background knowledge of an ML model's characteristics or training set tries to reproduce the model by "stealing it", in these types of attacks. ML-as-a-service services, unlike traditional learning theory environments, may accept limited feature vectors as inputs and provide trust values with predictions.

2.3.2 Evasion Attacks

Moreover, the adversary during the research process, can conduct an evasion attack against classification, resulting in an incorrect machine interpretation. In this case, the adversary's target is to misclassify some data in order to, for example, stay stealthy or imitate some favourable behaviour. In terms of network anomaly detection, an intrusion detection system (IDS) can be avoided by interpreting the attack payload in such a manner that the target of the content can read it, but the IDS cannot, amounting to a misclassification. As a result, the perpetrator will damage the targeted device without being detected by the IDS. Another target of the intruder may be to induce concept drift in the system, resulting in persistent system re-training and dramatically deteriorating its efficiency.

The primary aim of this type of adversarial machine learning is to reduce the performance of the classification process that is based on machine learning. For classification problems, this can be interpreted as increase in false positives, in false negatives, or in both. For clustering problems, the aim is generally to reduce accuracy.

- **False positives:** In classification problems, such as spam detection, where there are two states (spam or normal), the aim of an attacker may be to make the targeted system falsely label many normal data as falsified data. This would lead to the decision-making system miss crucial information.
- **False negatives:** Using the same example, if the attacker aims to increase the false negatives, then many falsified data will actually be labelled as legitimate.
- **Both false positives and false negatives:** Here, the attacker aims to reduce the overall confidence of the user in the decision-making process by letting falsified data go through and by filtering out legitimate data.
- **Clustering accuracy reduction:** Compared to classification, the accuracy of clustering is less straightforward to evaluate. Here, we include a general reduction of accuracy as the overall aim of the attacker of a clustering algorithm.

2.4 Trustworthiness of Data Input to Machine Learning Algorithms

"AI Is Only as Good as the Data You Feed It" is a well-known phrase in the AI community and, indeed, stands true, as it reflects this reality from a technical perspective. AI solutions, and especially the latest Deep Neural Network (DNN) setups, are very efficient in capturing patterns in data both in supervised and unsupervised ways. In this regard, an AI system which is instantiated with a specific training set inherits the intrinsic characteristics of the that data. Hence, if a biased training set (within a given context) is used, then the trained AI system will gain only a partial knowledge of the context for which it was trained for. This may result to a poor performance during the actual deployment of the system in practice. This is just an indication of the implications that may emerge due to the poor data quality.

However, apart from the quality of the data, the aim of this section is to highlight the importance of the trustworthiness of data which are being fed into the AI systems. Following the same mindset, we argue that "AI Is Only as Trustworthy as the Data You Feed It". In the context of adversarial machine learning and more specifically, in the context of poisoning and evasion attacks, the community has witnessed a series of events at stages of the machine learning pipeline (training and production) where attackers try to highjack the training process or to evade the inference process of AI systems. In both cases, the attackers inject small perturbations in the data which are just-enough in order to either lead to a faulty trained systems or to fool the system at the inference stage.

It becomes clear, that in order to safeguard AI systems we need, not only to enhance the robustness of the AI models per se, but also to deploy additional techniques that can guarantee the operational assurance of the components taking part in the data processing pipelines of AI systems. Thus, we argue that beneficial techniques, such as Adversarial Training or Defensive distillation [1], can be complemented event further by solutions that technically can offer verifiable evidence on the provenance and integrity of the data, and the legitimate operational state of the data generators. Especially, in the case of smart manufacturing, where multiple heterogenous devices support different production lines that generate diverse data flows, it is crucial to identify these roots of trust.

In the context of smart manufacturing, attestation can be used as a solution to guarantee the operational assurance of systems and to a certain extend to be used as the root of trust for the generated data flows.

Particularly, heterogeneous components must be enabled to make and prove statements about the integrity of their produced data so that other components can align their actions appropriately and an overall system state can be assessed. This goes substantially beyond simple authorization schemes telling who may access whom but will require understanding of semantics of requests and chains of effects throughout the system and an analysis both statically at design-time and dynamically during runtime.

2.4.1 Attestation for the Trustworthiness of Data Generators

Remote attestation is an efficient mechanism to provide evidence of the integrity status of a remote component. It is typically realized as a challenge-response protocol that allows a trusted party (verifier) to obtain an authentic and timely report about the state of an untrusted, and potentially compromised, remote device (prover). A prominent root of trust to enable attestation is the Trusted Platform Module (TPM). The TPM allows to implement remote attestation protocols in such a way that the anonymity of the platform is protected. Remote attestation services are currently used in a variety of privacy-preserving scenarios, ranging from attestation for isolated execution environments based on the -now outdated- Intel's Trusted Execution Technology [27], to more modern approaches used in conjunction with Intel's Software Guard Extensions, e.g. [28, 29].

From a high-level perspective, a remote attestation protocol requires that the prover creates an Attestation Key (AK) via the TPM, which is an asymmetric key pair used for signing quotes. A quote is a digitally signed report of the contents stored in selected Platform Configuration Registers (PCRs) of the TPM with the AK, i.e., a signature of the platform state. In order to preserve the anonymity, the prover has the ability to create as many AKs as they wish, but it is required that each AK be certified by a trusted third party called the Privacy Certification Authority (PCA). A verifier can trust the platform if it successfully verifies that a quote is a valid signature over expected PCR values with a certified AK.

The aforementioned process is the pilar in the trusted computing field in order to establish trust among different TPM-enabled entities. The benefits of this solution have led to the realisation of numerous attestation approaches, while several implementations and research endeavours have emerged with particular focus in IoT environments. More specifically, leveraging cryptographic techniques for protecting and proving the authenticity and integrity of computing platforms, and in turn, the data stemming from those platforms, has resulted to a rich scientific field. Both integrity and authenticity are two indispensable enablers of trust. Whereas integrity provides evidence about correctness, authenticity provides evidence of provenance.

Typical attestation solutions measure the load-time integrity of user-space applications and files read by the root user during runtime. This is the Binary-Based Attestation (BBA) scheme proposed by TCG, where measurements and attestation consider hashes of binaries. Other solutions, focus on the attestation of only a set of critical properties of the attested devices in order to provide more efficient and flexible schemes on the basis of Property-based Attestation (PBA) [30]. The aforementioned schemes offer a rather static assertion on the integrity of a platform and its configuration. To tackle this limitation, Control-flow Attestation (CFA) solutions suggest the acquisition of measurement that reflect the run-time behaviour of a processes in order to detect attacks that try to evade the legitimate execution behaviour of a system during runtime.

Considering the above, AI-enabled and IoT-based smart manufacturing industries can take advantage of remote attestation mechanisms in order to establish trust among all the components that operate collaboratively in a manufacturing process. By having indisputable evidence on the configuration and/or runtime integrity of shop floor devices, the cyber-attack surface is by far minimised leading and establishing trust among devices on the shop floor. More specifically, in order to guarantee the integrity and correctness of data, property-based attestation [30] seems to be the perfect fit. By identifying these exact properties that need to be attested on manufacturing systems, A PBA mechanism can guarantee the operational assurance of component which are responsible for data generation which are fed into the data pipeline of AI systems.

Attestation can ensure that the data sent from one device to another device has not been tampered, and this could be ensured in all data processing phases, i.e., during transport, during generation or processing on the originating device [31]. Attestation can be used as a provenance mechanism, as data exchanged between devices in a network can be authenticated along with a proof of integrity of all software involved in its generation and processing. The strategy used in [31] to achieve this, was to decompose the software of embedded devices into simple interacting modules reducing the amount and complexity of software that needs to be attested, i.e., only those modules that process the data are relevant.

In the context of AI-fuelled smart manufacturing, where the trustworthiness of data is a crucial requirement that needs to be met, remote attestation seems a viable solution to guarantee the integrity of data and minimize the possibility of adversarial attacks against AI systems.

2.5 Discussion and Critique

Cyber defense in the manufacturing industry is divided into two categories: static defense and active defense. Static defense methods are centered on adhering to common industrial rules and specifications. Cryptographic corrective actions, intrusion detection and prevention systems, human coaching, and incident response management are examples of dynamic defense mechanisms. Although static defense is a vital step toward improving overall security posture, it is relatively simple, so more specifics are overlooked. Manufacturing and smart manufacturing environments contain hundreds or even thousands of devices, the majority of which are Internet of Things (IoT) devices. Cryptographic primitives are well-known and broadly used in systems to ensure data confidentiality and integrity. The usage of symmetric encryption algorithms, public key infrastructure (PKI), hybrid encryption schemes, cryptographic hash functions, and digital signatures can secure the integrity of the data, can be used for authentication, ensures that a sender when sending a message, cannot deny the authenticity of a message that he sent to the recipient, non-repudiation and many other aspects of security. Another cyber defense mechanism is intrusion detection systems in smart manufacturing network-based environments which are categorized in *Host-based IDS* and *Knowledge-based IDS* [34, 35]. Host-based IDS gather data on single hosts compared to Knowledge-based IDS which are accumulating information about previous security flaws and find patterns to detect intrusions. Both of these security mechanisms work with signature-based security and basically, the limitation of signature-based security is that they cannot capture so easily zero-day exploits and newly introduced attacks. Due to the complexity of modern systems and smart IoT devices used in smart manufacturing environments traditional machine learning (tree based, Bayesian based, SVMs, etc,) systems and models are operating based on input data (e.g. Network data, images from robots, sound data, coordinates etc.) that is collected mainly on network endpoints which are monitored by our system. Based on this input they can perform a number of decisions (e.g. Alert the system administrator, raise an incident etc.) based on classification models which, however, can be considered as limited (Zhang *et al.* (2019) [25], Banerjee *et al.* (2018) [4] and Meng Qu *et al.* (2018) [32]), because they do not take advantage of enhanced understanding of events that may happen in other parts of the network, as well as the luck of appropriateness for aggregating heterogeneous neighbours with different content features. By features we refer to the features extracted from monitoring and processing collected network and host-based data that can be used in the classification of specific attack vectors. More specifically, there is no correlation of data acquired by different individual sources. In the industrial sector, and even in the scientific literature, for example, deep learning has been largely applied to datasets in which the training data are: (i) independent of each other, and (ii) homogeneous, i.e., the subjects of the classification or regression are instances with same entity type, whereby each section in the schematic diagram has a consistent interpretation and format. Thus, there is a need to develop more accurate classification models when it comes to detecting a wider range of attacks, based on the classification of malicious and benign network traffic, in collaboration with advanced AI.

2.6 Outlook – Road Ahead

Entities in smart manufacturing infrastructure are most probably heterogeneous and endowed with characteristics that change dynamically over time compared to their subsequent interactions. To apply deep learning to such entities, for example, for classification, one must first assimilate the encounters into the feature engineering process in a structured manner. The reason for these research questions is to demonstrate how, in this regard, the present state of graph machine learning is insufficient and needs supplementation with a rigorous function engineering framework in space and time. Zhang *et al.* (2019) [25] provides enough proof to challenge the concept that traditional machine learning methods are not suitable to create the most complete and concrete classification model. Also, in the H2020

STAR project the approach that is investigated to overcome such challenges and limitations is using Graph machine learning and LSTM which shows promising results nonetheless there are still a number of open challenges to consider especially related to the order of monitored events and the time they are present in the system. We use this base and state of the art machine learning methods to challenge the most dangerous threat that is present to traditional machine learning classification models, the "Concept Drift" attack. In smart manufacturing "concept drift" attacks can apply in multiple examples, one of the examples is the temperature of a very critical room where IoT sensors are present. The attacker can manipulate the classification model changing its perspective by increasing very slowly the temperature of the room thus, impacting the manufacturing environment and causing huge damage to the machines. Graph machine learning is enhancing the knowledge, given to the classifier, by using different types of data produced by neighbouring endpoints, as well as the interaction of the neighbours with other devices and endpoints (entities). This is the difference between the intrinsic and extrinsic features based on which the classification takes place. Each of these objects has properties, i.e., characteristics, that are inherent to them. It should be noted that these intrinsic properties are often transient and therefore necessitate a sequential treatment. Extrinsic characteristics, on the other hand, emerge from the entities' relations with one another, and are influenced by different environmental parameters. When entities communicate, their extrinsic properties, both of which are dynamic, must be modified to accommodate the changing probability that any particular entity bears. The combination of both intrinsic and extrinsic features enhances the knowledge of the classifier and this is the benefit that Graph machine learning offers to other traditional machine learning methods. A more specific and novel solution to the above procedure is the usage of Bipartite Graphs for hypergraph machine learning. The solution requires a combination of Bipartite graph models with advanced AI LSTM (Long Short-Term Memory) agents. LSTMs, introduced by Hochreiter *et al.* [33], and their ability to learn on data with relationships and with long-range temporal dependencies, makes them a well-suited technology for phenomena with spatial and time characteristics such as time series prediction, machine translation, speech recognition, language processing. Since there can be unexplained lags between significant events, LSTM is useful for sorting, classifying, and drawing conclusions based on time series data. The reason for using this specific type of ML-based agents is the fact that they take into account the time dependency which is crucial in cybersecurity attacks. Based on the use of LSTMs a new classification framework can be designed to prove the efficiency and effectiveness in accuracy compared to traditional classifiers. The most usual problem which has been indicated from traditional AI methods is the inability of the methods to successfully flag "Concept Drift" type of attacks. In these types of attacks, the attackers manipulate the data slightly as the time goes which can disarms the ability of the traditional AI methods to successfully classify an attack. Thus, the use of LSTM is imperative for creating the right framework.

2.7 Conclusions

This chapter focused on the AI adversarial tactics against smart manufacturing in order to identify the gaps that enable cyber attackers to manipulate AI systems. As such systems have become an integral part of the modern production lines for supporting a wide range of operations, from predictive maintenance to safe human-robot collaboration, among others, such systems have attracted the interest of attackers. In this direction, this chapter offered a review on the current status of smart manufacturing domain by highlighting the emerging threats and it overall security posture. In this context, we elaborated on the emerging threats of poisoning and evasion attacks against AI manufacturing systems and how attestation mechanism can be used to guarantee the trustworthiness of generated data in the manufacturing domain. This analysis led to a discussion on the road ahead that gave the chance to document the benefits of including Graph machine learning and LSTM for building robust AI setups for smart manufacturing.

Acknowledgements

This work has been carried out in the H2020 STAR project, which has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 956573.

References

[1] Pitropakis, N., Panaousis, E., Giannetsos, T., Anastasiadis, E. and Loukas, G. (2019). A taxonomy and survey of attacks against machine learning. Computer Science Review, 34, 100199.

[2] Loukas, G., Karapistoli, E., Panaousis, E., Sarigiannidis, P., Bezemskij, A. and Vuong, T. (2019). A taxonomy and survey of cyber-physical intrusion detection approaches for vehicles, Ad Hoc Netw. 84, 124–147.

[3] Rouani, B.D., Samragh, M., Javidi, T. and Koushanfar, F. (2019). Safe machine learning and defeating adversarial attacks, IEEE Secur. Priv. 17(2), 31–38.

[4] Banerjee, N., Giannetsos, T., Panaousis, E. and Took, C.C. (2018). "Unsupervised Learning for Trustworthy IoT", In IEEE International Conference on Fuzzy Systems (FUZZ-IEEE).

[5] Hasan, T., Akhunzada, A., Giannetsos, T. and Malik, J. "Orchestrating SDN Control Plane towards Enhanced IoT Security", In Proceedings of 2020 IEEE Conference on Network Softwarization.

[6] Gisdakis, S., Giannetsos, T. and Papadimitratos, P. (2015). "SHIELD: A Data Verification Framework for Participatory Systems", In Proceedings of the 8th Conference on Security and Privacy in Wireless and Mobile Networks.

[7] Xu, L.D., Xu, E.L. and Li, L. (2018). Industry 4.0: state of the art and future trends. Int J Prod Res; 56(8):2941–62.

[8] Jung, K., Morris, K.C., Lyons, K.W., Leong, S. and Cho, H. (2015). Mapping Strategic Goals and Operational Performance Metrics for Smart Manufacturing Systems. Procedia Computer Science.

[9] Schönsleben, P., Fantana, F. and Duchi, A. (2017). What benefits do initiatives such as industry 4.0 offer for production locations in high-wage countries? CIRP 50th Conference on Manufacturing Systems.

[10] Li, D., Tang, H., Wang, S.Y. and Liu, C.L. (2017). A big data enabled load-balancing control for smart manufacturing of Industry 4.0. Cluster Comput. J. Netw. Softw. Tools Appl. 20(2), http://dx.doi.org/10.1007/s10586-017-0852-1.

[11] Mueller, E., Chen, X.L. and Riedel, R. (2017). Challenges and requirements for the application of Industry 4.0: a special insight with the usage of cyber-physical system. Chin. J. Mech. Eng. 30(5), http://dx.doi.org/10.1007/s10033-017-0164-7, 9.

[12] Kusiak, A. (2016a). "Put Innovation Science at the Heart of Discovery." Nature 530(7590): 255–255.

[13] Mittal, S., Khan, M.A., Romero, D. and Wuest, T. (2017). Smart manufacturing: Characteristics, technologies and enabling factors. Proceedings of the Institution of Mechanical Engineers, Part B: Journal of Engineering Manufacture.

[14] ISO 10218-1 2008 standard (2008). Robots for industrial environments—safety requirements, part 1: robot.

[15] Cerrudo, C. and Apa, L. (2017). "Hacking Robots Before Skynet". In Cybersecurity Insight, IOActive Report, Seattle, USA.

[16] Radvanovsky, B. and Brodsky, J. (2015). Project SHINE (SHodan INtelligence Extraction), Findings Report.

[17] Tuptuk, N. and Hailes, S. (2018). Security of smart manufacturing systems. Journal of manufacturing systems, 47, 93–106.

[18] Nilufer Tuptuk and Stephen Hailes, The cyberattack on Ukraine's power grid is a warning of what's to come, https://theconversation.com/the-cyberattack-on-ukraines-power-grid-is-a-warning-of-whats-to-come-52832

[19] Antonakakis, M., April, T., Bailey, M., Bernhard, M., Bursztein, E., Cochran, J., *et al.* (2017). Understanding the Mirai Botnet 26th USENIX Security Symposium (USENIX Security 17), USENIX Association, Vancouver.

[20] Sturm, L.D., Williams, C.B., Camelio, J.A., White, J. and Parker, R. (2014). Cyber-physical Vunerabilities In Additive manufacturing systems, in international solid freeform fabrication symposium proceedings, pp. 951–963.

[21] Kim, A.C., Park, W.H. and Lee, D.H. (2013). A study on the live forensic techniques for anomaly detection in user terminals. International Journal of Security and Its Applications, 7(1), 181–187.

[22] Chandola, V., Banerjee, A. and Kumar, V. (2009). Anomaly detection. ACM Computing Surveys, 41(3), 1–58.

[23] Garcia, R.F., Rolle, J.L.C. and Castelo, J.P. (2011). A review of SCADA anomaly detection systems. Advances in Intelligent and Soft Computing, 87, 405–414.

[24] Rabatel, J., Bringay, S. and Poncelet, P. (2011). Anomaly detection in monitoring sensor data for preventive maintenance.Expert Systems with Applications, 38, 7003–7015.

[25] Zhang, Jiliang and Li, Chen. (2019). Adversarial Examples: Opportunities and Challenges. In IEEE Transactions on Neural Networks and Learning Systems.

[26] Behzadan, Vahid and Munir, Arslan. (2017). Vulnerability of Deep Reinforcement Learning to Policy Induction Attacks. arXiv:1701.04143.

[27] Goldman, Ken: IBM's Software TPM 2.0 and TSS, https://sourceforge.net/projects/ibmswtpm2/, https://sourceforge.net/projects/ibmtpm20tss

[28] Ibrahim, F.A. and Hemayed, E.E. (2019). Trusted cloud computing architectures for infrastructure as a service: Survey and systematic literature review. Computers & Security 82, 196(226).

[29] TCG: TCG Guidance for Securing Network Equipment Using TCG Technology Version 1.0 Revision 29 (jan 2018), https://trustedcomputinggroup.org/wp-content/uploads/TCG_Guidance_for_Securing_NetEq_1_0r29.pdf

[30] Koutroumpouchos, N., Ntantogian, C., Menesidou, S.A., Liang, K., Gouvas, P., Xenakis, C. and Giannetsos, T. (2019, June). Secure edge computing with lightweight control-flow property-based attestation. In 2019 IEEE Conference on Network Softwarization (NetSoft) (pp. 84–92). IEEE. DOI: 10.1109/NETSOFT.2019.8806658

[31] Abera, T., Bahmani, R., Brasser, F., Ibrahim, A., Sadeghi, A.R. and Schunter, M. (2019, January). DIAT: Data Integrity Attestation for Resilient Collaboration of Autonomous Systems. In NDSS.

[32] Meng Qu, Jian Tang and Jiawei Han. (2018). Curriculum Learning for Heterogeneous Star Network Embedding via Deep Reinforcement Learning. In WSDM. 468–476.

[33] Hochreiter, Sepp and Schmidhuber, Jürgen. (Nov. 1997). "Long Short-Term Memory". In: Neural Computation 9.8.

[34] Papamartzivanos, D., Mármol, F.G. and Kambourakis, G. (2018). Dendron: Genetic trees driven rule induction for network intrusion detection systems. Future Generation Computer Systems, 79, 558–574. DOI: 10.1016/j.future.2017.09.056

[35] Papamartzivanos, D., Mármol, F.G. and Kambourakis, G. (2019). Introducing deep learning self-adaptive misuse network intrusion detection systems. IEEE Access, 7, 13546–13560. DOI: 10.1109/ACCESS.2019.2893871

DOI: 10.1561/9781680838770.ch3

Chapter 3

Knowledge Modelling and Active Learning in Manufacturing

By Jože M. Rožanec, Inna Novalija, Patrik Zajec, Klemen Kenda and Dunja Mladenić

The increasing digitalization of the manufacturing domain requires adequate knowledge modeling to capture relevant information. Ontologies and Knowledge Graphs provide means to model and relate a wide range of concepts, problems, and configurations. Both can be used to generate new knowledge through deductive inference and identify missing knowledge. While digitalization increases the amount of data available, much data is not labeled and cannot be directly used to train supervised machine learning models. Active learning can be used to identify the most informative data instances for which to obtain users' feedback, reduce friction, and maximize knowledge acquisition. By combining semantic technologies and active learning, multiple use cases in manufacturing domain can be addressed taking advantage of the available knowledge and data.

3.1 Introduction

Digitalization enables collecting and storing data in a digital format. Digital data enables changes at the process, organization, and business domain levels [1]. In manufacturing, it allows to achieve increased process efficiency (with lower performance variability and less unplanned downtime), a more efficient use of resources (e.g., lower energy consumption), increased safety and sustainability, product quality, and reduced product launch time [2–4]. Smart factories are built based on three principles [2]: cultivate digital people, introduce agile processes, and configure modular technologies. [5] identifies four digitalization challenges in manufacturing: how to digitally augment human work, enable worker-centric knowledge sharing, create self-learning manufacturing workplaces, and enable mobile learning. These challenges and benefits were recognized by several national and international initiatives (Advanced Manufacturing (USA), Industry 4.0 (Germany and the European Union) [6], Made in China 2025, New Robot Strategy (Japan), New Industrial France, High-Value Manufacturing (UK), Make it Happen (Australia)), and new paradigms created to realize them. Among such paradigms, we find Cyber-Physical Systems [7], Digital Shadows, and Digital Twins [8]. Cyber-Physical Systems were conceived as smart and embedded systems that result from the integration of physical and computational processes [9, 10]. In Digital Shadows, the data flow is unidirectional (from the physical counterpart to the digital replica), while in Digital Twins, this flow is bidirectional (changes in the digital object can lead to changes in the physical object) [11]. Multiple authors proposed enhancing the Digital Twins providing cognitive capabilities using a knowledge graph [12, 13]. Such technologies, along with the Internet of Things and Artificial Intelligence, bring added value into industrial value chains [14].

To capture data in a digital form, sensors and software, such as as Enterprise Resource Planning (ERP) or Manufacturing Execution Systems (MES), are used. There are, however, many operational aspects and contextual information the employees are aware of that the sensors and the software systems mentioned above do not capture. Thus, it is essential to develop interfaces and mechanisms to gather such information while minimizing interaction friction with end-users. Some examples can be found in other domains, where to mitigate the knowledge gap, researchers developed conversational interfaces that identify missing knowledge and ask the users to provide it [15, 16]. In such a context, semantic technologies and active learning can play a crucial role. Semantic technologies enable encoding domain knowledge (in ontologies and knowledge graphs) and provide means to perform inference (considering rules and logics) [17]. On the other side, active learning allows one to choose the most informative pieces of data to gather additional insights from experts.

This chapter discusses semantic knowledge representations (ontologies and knowledge graphs), the usage of active learning, and use cases in the industrial domain that benefit from both.

3.2 Semantic Knowledge Representations

3.2.1 Ontologies in Manufacturing

Ontologies are explicit specifications of a conceptualization (an abstract, simplified view of the world) regarding objects, concepts, and entities, and the relationships between them [18]. One of the main issues regarding knowledge management in the manufacturing domain is the wide range of concepts, problems, and configurations present [19]. A possible solution to this is the usage of semantic technologies. Ontologies provide a formal specification of a shared conceptualization in the domain of interest by defining concept hierarchies, taxonomies, and topologies [20, 21]. They provide information interoperability between different domains and enable reasoning. Among the ontology use cases in manufacturing mentioned in the literature, we find knowledge sharing and reuse in distributed manufacturing settings [22], linking between product assemblies and manufacturing resources using manufacturing operations [23], and production line processes [24]. Several ontologies were considered and developed in the manufacturing domain. Upper ontologies provide high-level concepts that can be extended to create domain-specific ontologies. Among the upper ontologies we find the Basic Formal Ontology (BFO) [25], Suggested Upper Merged Ontology (SUMO) [26], Descriptive Ontology for Linguistic and Cognitive Engineering (DOLCE) [27], General Formal Ontology (GFO) [28], Object-Centered High-level Reference Ontology (OCHRE) [29], Politecnico di Milano–Production Systems Ontology (P-PSO) [30], Manufacturing Reference Ontology (MRO) [31], Manufacturing Systems Ontology (MSO) [32], and the Manufacturing System Engineering (MSE) ontology model [33].

The BFO ontology attempts to model time and space. To that end, it divides entities into two disjoint categories: continuants (something that exists at a point in time) and occurrents (something that is realized in time, e.g., processes and events). SUMO is considered the largest formal public ontology, mapping the whole WordNet lexicon. It divides entities into two disjoint categories: physical (represents objects and processes) and abstract (represents sets, propositions, quantities, and attributes). The upper ontology is complemented with the MId-Level Ontology (MILO), and domain-specific ontologies are developed on top of them. DOLCE was developed to capture ontological categories underlying natural language and human commonsense. Their focus is to describe categories as cognitive artifacts as represented in human perception, rather than in the intrinsic nature of the world. The ontology divides entities into two categories: *endurants* (continuants) and *perdurants* (occurrents). A mapping between BFO and DOLCE was proposed in [34]. GFO provides a different worldview, dividing entities into two categories: presential and process. *Presentials* refer to entities that are entirely present at a given point in time. To model how *presentials* acquire different values in time but remain the same entity, they refer to *persistents* (a specific universal representing the *presentials*). The processes represent functions that have a temporal extension and cannot be wholly present only at a given point in time. The persistent-presential aspects are discussed in OCHRE under the terms of *thick* and *thin objects*, where the thick objects refer to aspects that change over time, while the thin object refers to core aspects that remain the same through time. While the ontology distinguishes between *endurants* and *perdurants*, it does so by modeling participation as a special case of parthood and avoids assuming two separate domains. P-PSO was designed as a meta-model to describe the manufacturing domain from an object-oriented perspective. When doing so, it considers three aspects of the manufacturing setting: physical (entities' material definition), technological (system functional view), and control (production operation procedures) aspect. The MSO evolves the P-PSO, addressing a wider domain, built with a different purpose, and providing a different approach to the control and visualization aspects. Regarding the domain, high-level classes are defined to address all types of industry, and specific classes are defined as specializations of such high-level objects. In particular, the ontology extends the scope to logistics and the process industry. In contrast to the P-PSO, which provides a general taxonomy but does not define a specific usage, MSO was designed for production system control. Finally, regarding control and visualization aspects, P-PSO defines entities and relationships to be considered for manufacturing control. In contrast, MSO provides definitions at a conceptual level, assuming the ontology only interacts with different software placing its interest on the outcomes, without the need to represent the inner design and working of the software service. The MRO was designed as an upper manufacturing ontology, defining the terminology based on existing standards. A different approach was adopted by the MSE ontology model, which provides a model to support information autonomy and facilitate information exchange between inter-disciplinary engineering design teams while leaving to each team freedom to adopt their terminology. The aforementioned upper ontologies are considered when creating domain-specific ontologies.

Manufacturing always relates to a specific product, and authors developed specific ontologies to describe them. [35] developed PRONTO (PRoduct ONTOlogy), which defines concepts, relationships, and axioms mainly related to the manufactured products' structure. The ontology considers raw materials, how those are assembled into a product, and derivative products. [36] noticed that ontologies and standards aim to facilitate a common grounding by sharing expert knowledge and finding agreement on a particular domain. They developed the ONTO-PDM ontology based on existing models[1] from the IEC 62264 standard.

Products cannot be developed without a specific manufacturing process. [37] developed the Process Specification Language (PSL) ontology to describe manufacturing processes throughout the manufacturing life cycle. [24] developed and applied a meta-model to describe a material-processing production line, which supports defining the behavior of an entity over time through state transitions. [38] developed the Manufacturing Resource Capability Ontology (MaRCO), to describe the capabilities of manufacturing resources, concentrating solely on machines and tools, so that can be used to support semi-automatic system design and auto-configuration of production systems. Another effort to describe products, production processes and resources is the P2 ontology, developed by [39]. [40] describe a manufacturing ontology-based on the DOLCE ontology and the Adaptive Holonic Control Architecture for distributed manufacturing systems (ADACOR) [41], that describes manufacturing scheduling and control operations. Another view on scheduling was developed by [42], who introduced the SIMPM (Semantically Integrated Manufacturing Planning Model) ontology, modeling manufacturing planning task according to time, variety, and aggregation. [43] presents the Supply Chain Operations (SCOR) ontology to facilitate the interoperation between applications involved in the supply chain. A product data model in a cloud manufacturing context was developed by [44]. Finally, additive manufacturing was subject of several ontologies [45, 46].

Another relevant aspect to the manufacturing domain is the sensors, which enable data gathering. OntoSensor [47] aims to provide a broad knowledge base of sensors for query and inference, based on the SensorML standard.[2] In the same line, [48] proposed an ontology to describe sensors' capabilities and operations. [49] developed WISNO (Wireless Sensor Networks Ontology) to deduce high-level information from low-level, implicit context and checking ontologies' consistency. [50] describes an ontology to characterize sensor capabilities and properties as the composition of their building blocks through three description levels: domain concepts, abstract sensor properties, and concrete properties.

[1] IEC 62264 models are: Product Definition, Material, Equipment, Personnel, Process Segment, Production Schedule, Production Capability, and Production Performance.

[2] SensorML is an approved Open Geospatial Consortium standard. More details are available at https://www.ogc.org/standards/sensorml

3.2.2 Methodologies for Ontology Design in Manufacturing

Rarely an ontology satisfies all the requirements and frequently needs to be extended, or new ontologies need to be developed to cover a new domain. Multiple methodologies were described in the literature to build an ontology. While each methodology has a different emphasis, there is consensus on most steps to be followed:

- identify the problem to be solved, opportunity areas, and a potential solution [51–53]
- decide on the formality level required [52]
- define the problem, scope and competency questions [54]
- elicit required knowledge from multiple sources. Identify key concepts and relationships. Identify terms that refer to the concepts and relationships. [51–53]. When doing so, consider the MIREOT guidelines [55].
- evaluate against a frame of reference [51, 53]

Specific methodologies were developed to guide the ontologies' construction in the manufacturing domain. [56] proposed a six-stage methodology: identify root concepts of taxonomies, identify existing taxonomies, create taxonomies, application test, build terms thesaurus, and refine the integrated taxonomy. [57] also defined a six-step methodology but provided a different procedure: specification (determine scope and granularity), conceptualization (acquire knowledge), formalization (structure acquired knowledge), population (convert acquired knowledge into frame-based representation), evaluation (validate accuracy and completeness), and maintenance (update the ontology once established). A different approach was developed by [22], who combined the Unified Modelling Language and the Object Constrained Language to translate entities from a software object model to ontology entities. [58] developed a four-step methodology using a Simple Knowledge Organization System (SKOS) framework to develop a thesaurus of concepts, to then identify relevant classes and provide logical constraints and rules. [59] suggests a three-step methodology, inspired in [60], that requires an ontology requirements specification (purpose, scope, and ontology requirements analysis), an analysis of existing resources (reuse ontological and non-ontological resources), and a conceptualization and formalization. A similar approach was developed by [61], with an emphasis on manufacturing design. [62] defined a nine-step methodology for developing process ontologies. First, it requires defining the project's purpose and scope, identifying potential classes and formal attributes, and writing them down to a context table. Concepts and subconcepts should be drafted to a lattice, which is used to resolve inconsistencies, and then converted to a class hierarchy. The last steps correspond to integrate the hierarchy with some upper ontology and the classes formally defined through axioms and relationships. Finally, [63] envisions a different scenario, creating an ontology building methodology for Cyber-Physical Systems in the manufacturing domain. The methodology consists of three steps: ontology requirements specification (based on project requirements), lightweight ontology building (considering requirements, information resources, and other lightweight ontologies), and heavy-weight ontology building (taking into account the lightweight ontology and ontology design patterns).

3.2.3 Knowledge Graphs in Manufacturing

Among the rich literature describing knowledge graphs, there is no agreed unique definition for them. The knowledge graphs are built upon the idea that graphs can be used to capture knowledge. Nodes are used to define abstractions and instantiate entities, which can be linked with edges, representing relationships [64]. They can be either domain-specific or domain independent [65]. Many implementations constrain the edges in knowledge graphs according to some schema or ontology [66], providing a formal concepts' definition. [67] provides a comprehensive introduction to knowledge graphs, discussing data models, schemas, deductive and inductive techniques, quality dimensions, refinement methods, and prominent open and enterprise knowledge graphs. Deductive inference can be used to derive new knowledge from existing data and rules known *a priori*. Inductive knowledge, on the other side, is acquired by generalizing patterns from input observations, either using supervised or unsupervised methods. Knowledge graph refinement attempts to identify wrong information in the graph (ensure that it is free of error) and complete missing information (satisfy completeness) [65]. Such tasks can benefit from knowledge graph embeddings, which reduce nodes and edges to continuous vector spaces while preserving the inherent graph structure [68]. When assessing the quality of a knowledge graph, [67] highlights four quality dimensions: accuracy (the extent to which the knowledge graph represents the real-world domain), coverage (avoid the omission of elements that are relevant to the specific domain), coherency (conformity to formal schema or ontology), and succinctness (avoid irrelevant data). [69] describes a wide range of quality metrics, classifying them in four quality categories described by [70]: intrinsic data quality (quality of data on its right, regardless of the use case), contextual data quality (assessed concerning the task at hand), representational data quality (relates to the format and meaning of the data), and accessibility data quality (relates to how data can be accessed, considering accessibility, licenses, and interlinking).

Knowledge graph implementations can adopt one of three assumptions: open world, locally closed world [71], or closed world assumption. Open world assumption considers that a statement can be true irrespectively of whether it is known to be true since there is much unknown information compared to the encoded knowledge. The local-closed world assumption considers the knowledge representation is locally complete. The truth regarding a statement can be determined as long as the set of existing object values for a subject and predicate are not empty. Finally, the closed-world assumption assumes that only statements known to be true can be true.

Manufacturing knowledge is gaining an increasing amount of attention [72]. The use of knowledge graphs to model it was reported in multiple scenarios. [73] describe building and using a knowledge graph to integrate information of products and equipment obtained from heterogeneous data sources. The knowledge graph is a cornerstone to an intelligent manufacturing equipment information system. [74] report encoding purchase records data in a supply chain knowledge graph and use embeddings to recommend the best suppliers for the purchase demand. [75] use natural language processing to extract disassembly data (entities and the nature of components) and then encode it in a knowledge graph, which helps to acquire, analyze and manage disassembly knowledge. A different purpose is envisioned by [76], who integrate semantic information of the workers with temporal profiling information, and facial recognition. Finally, [77, 78] describe a knowledge graph to integrate information regarding Industry 4.0 standards and standardization frameworks. Using graph embeddings, they can detect standards relatedness, identify similar standards and unknown relations.

3.3 Active Learning

Active learning is a field of machine learning that studies how to select unlabeled data samples and query an information source to label the selected samples [79–81]. The underlying assumption is that unlabeled data is abundant, and labeling resources are scarce. Therefore, it is necessary to devise mechanisms that enable the identification and selection of samples with a higher information potential. The promise to reduce the amount of data required to train new models has driven increased interest in active learning in the academic community. At the same time, the adoption remains low in industry [82].

Three different active learning scenarios are described in the literature [83]: membership query synthesis, stream-based selective sampling, and pool-based active learning. In membership query synthesis [84, 85] an algorithm creates its instances (queries) from an underlying distribution to ask the expert if the instance corresponds to a particular label. Stream-based selective sampling considers one unlabeled instance at a time, evaluating its informativeness against the query parameters. The learner decides whether to query the teacher or assign the label by itself. Finally, in pool-based sampling, unlabeled instances are drawn from the entire data pool and assigned an informative score. Most informative instances are selected, and their labels requested. While most methods rely on model uncertainty and clustering to choose the unlabeled examples [86, 87], new approaches were developed based on adversarial sampling, Bayesian methods, and weak supervision. [88] introduced a new approach to active learning (Generative Adversarial Active Learning (GAAL)) by leveraging Generative Adversarial Networks (GAN). The purpose of the GANs is to generate informative instances based on a random sample of unlabeled instances close to the decision boundary. [89] evolved this concept generating synthetic data with a conditional GAN, which learns to create a specific instance leveraging additional data regarding the desired target label. [90] introduced the variational adversarial active learning, sampling instances using an adversarially trained discriminator to predict whether the instance is labeled or not based on the latent space of the variational auto-encoder. Since the sampling ignores the instance labels, the discriminator can end up selecting instances that correspond to the same class, regardless of the proportion of labeled samples of such class. To solve such an issue, [91] developed a semi-supervised minimax entropy-based active learning algorithm that leverages uncertainty and diversity in an adversarial manner. Another approach was developed by [92], who, instead of uncertainty sampling, used a GAN to generate high entropy samples and retrieve similar unlabeled samples from available data to acquire the corresponding labels. A variation to GAAL, and based on previous work by [93], [94] developed a Bayesian generative active deep learning approach, performing a joint training of the generator (a variational autoencoder) and the learner, which requires smaller sample sizes and a single training stage. Different approaches were developed by [95–97], who explored using active learning in a weak supervision setting.

Despite the wide range of active learning approaches, there is currently a research void regarding the use of active learning in the manufacturing domain [98]. It was successfully applied to predict the local displacement between two layers on a chip in the semi-conductor industry [99], for automatic optical inspection of printed circuit boards [100], to improve the predictive modeling for shape control of composite fuselage [101], and in multi-objective optimization [102].

3.4 Use Cases and Open Challenges

Semantic technologies and active learning can be used to identify missing knowledge. Within the semantic domain, [103] proposed a typology of missing knowledge, identifying three types of missing knowledge: *abstraction dimension* (how the knowledge is contained inside the KG structure), *terminological knowledge* (how to map terms to concepts), and *question-answering dimension* (how the lack of knowledge affects the answering process). [104] proposed to frame the missing knowledge problem as an anomaly detection problem, where they use a heuristic to identify missing knowledge in system rule bases. They take into account user input during the inference process for items that are considered *askable*. [105] suggested an approach based on first-order logic and dual polynomials. They use triples consisting of a question, answer, and a label that can indicate if the answer is missing or wrong. For missing answers, they developed heuristics to create potential answers that comply with a closed world setting. [106] proposed developing an interface to issue SQL queries that can target either a relational database or crowdsource certain operations, such as find new data or perform non-trivial comparisons. The authors consider that while many operations can be successfully completed with data within a database, humans can assist with operations such as gathering missing data from external sources, moving towards an open-world assumption.

[107] combined ontologies with natural language processing to develop a question answering interface that enabled users to access available underlying data sources. Among other results, the authors highlighted how such a system provided a positive experience to the users, doubling user retention. [68] describes the use case, using a knowledge graph to simplify question answering by organizing them in a structured format. [108] tackles question answering by creating vector embeddings of questions and knowledge graph triples so that the question vectors end up close to the answer vectors. A different approach was considered by [16], who developed *Curious cat*. This application leverages a semantic knowledge base and user's contextual data for knowledge acquisition through question-answering. A similar knowledge acquisition approach for the manufacturing domain was envisioned by [109], who developed an ontology to model user feedback based on a given forecast and provided explanations. Following the need to augment human work with digital technologies and provide personalized information at the shop-floor level [110, 111] developed a smart assistant for manufacturing. The smart assistant creates directive explanations for the users by using heuristics and domain knowledge. The application tracks user's implicit and explicit feedback regarding local forecast explanations, enabling application-grounded evaluations. Though the authors tested their approach on the demand forecasting use case, the application can be extended to other use cases. Other relevant use cases are the usage of semantic technologies to build a decision support system [112], automatically identify opportunities to enhance production scenarios [113], and intelligent condition monitoring of manufacturing tasks [114].

Though little research reports on the usage of active learning in manufacturing [98], we consider it can be widely applied in this domain. By selecting the most valuable instances to the system, it helps to minimize friction towards the end-user and collect valuable data [115]. Active learning can also increase the diversity of recommendations [116]. This approach can be used in applications recommending decision-making options to balance usual recommendations and decision-making options requiring more user feedback (more labeled instances) to enhance the underlying recommender system. Other relevant active learning use cases to manufacturing can be anomaly and outlier detection [117–119].

In the European Horizon 2020 project STAR (Safe and Trusted Human Centric Artificial Intelligence in Future Manufacturing Lines), knowledge modeling and active learning are used to gather locally observed collective knowledge regarding operations in the manufacturing lines and provide accurate context, relevant data, and decision-making options to the users. Among relevant use cases to the project are production planning (to gather additional context regarding downtimes and anomalies in production), optical quality inspection (to learn from images of defective parts), and logistics (to learn logisticians' decision-making based on available options).

3.5 Conclusion

Semantic technologies provide means to encode domain knowledge and enable deductive inference through reasoning engines. In this work we presented many upper-level and domain-specific ontologies from the manufacturing domain, and upon which new ontologies can be built. We also described multiple methodologies used to guide the ontology creation process, some of them specific to the manufacturing domain.

Semantic technologies can be leveraged for knowledge acquisition. Missing knowledge detection can be linked to a question-answering interface to gather required knowledge from the users. Similarly, active learning can be used can identify the most informative data instances and ask the users for feedback. This enables to gradually increase the dataset and its information density, which can be leveraged to train machine learning models, and enhance their performance. While little scientific literature reports on the usage of active learning in the manufacturing domain, multiple use cases can benefit from it, such as anomaly detection in production planning, optical quality inspection, and the recommendation of decision-making options.

Acknowledgements

This work has been carried out in the H2020 STAR project, which has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 956573.

References

[1] P. Parviainen, M. Tihinen, J. Kääriäinen, and S. Teppola, "Tackling the digitalization challenge: how to benefit from digitalization in practice," *International journal of information systems and project management*, vol. 5, no. 1, pp. 63–77, 2017.

[2] D. R. Sjödin, V. Parida, M. Leksell, and A. Petrovic, "Smart factory implementation and process innovation: A preliminary maturity model for leveraging digitalization in manufacturing moving to smart factories presents specific challenges that can be addressed through a structured approach focused on people, processes, and technologies," *Research-Technology Management*, vol. 61, no. 5, pp. 22–31, 2018.

[3] L. S. Dalenogare, G. B. Benitez, N. F. Ayala, and A. G. Frank, "The expected contribution of industry 4.0 technologies for industrial performance," *International Journal of Production Economics*, vol. 204, pp. 383–394, 2018.

[4] M. Demartini, S. Evans, and F. Tonelli, "Digitalization technologies for industrial sustainability," *Procedia manufacturing*, vol. 33, pp. 264–271, 2019.

[5] A. Richter, S. Vodanovich, M. Steinhüser, and L. Hannola, "It on the shop floor-challenges of the digitalization of manufacturing companies," 2017.

[6] F. Yang and S. Gu, "Industry 4.0, a revolution that requires technology and national strategies," *Complex & Intelligent Systems*, pp. 1–15, 2021.

[7] R. Rajkumar, I. Lee, L. Sha, and J. Stankovic, "Cyber-physical systems: the next computing revolution," in *Design automation conference*, pp. 731–736, IEEE, 2010.

[8] M. Grieves, "Digital twin: manufacturing excellence through virtual factory replication," *White paper*, vol. 1, pp. 1–7, 2014.

[9] H. J. La and S. D. Kim, "A service-based approach to designing cyber physical systems," in *2010 IEEE/ACIS 9th International Conference on Computer and Information Science*, pp. 895–900, IEEE, 2010.

[10] E. Negri, L. Fumagalli, and M. Macchi, "A review of the roles of digital twin in cps-based production systems," *Procedia Manufacturing*, vol. 11, pp. 939–948, 2017.

[11] W. Kritzinger, M. Karner, G. Traar, J. Henjes, and W. Sihn, "Digital twin in manufacturing: A categorical literature review and classification," *IFAC-PapersOnLine*, vol. 51, no. 11, pp. 1016–1022, 2018.

[12] J. Lu, X. Zheng, A. Gharaei, K. Kalaboukas, and D. Kiritsis, "Cognitive twins for supporting decision-makings of internet of things systems," in *Proceedings of 5th International Conference on the Industry 4.0 Model for Advanced Manufacturing*, pp. 105–115, Springer, 2020.

[13] J. M. Rožanec, J. Lu, J. Rupnik, M. Škrjanc, D. Mladenić, B. Fortuna, X. Zheng, and D. Kiritsis, "Actionable cognitive twins for decision making in manufacturing," *arXiv preprint arXiv:2103.12854*, 2021.

[14] I. Grangel-González, *A knowledge graph based integration approach for industry 4.0*. PhD thesis, Universitäts-und Landesbibliothek Bonn, 2019.

[15] A. Preece, W. Webberley, D. Braines, N. Hu, T. La Porta, E. Zaroukian, and J. Bakdash, "Sherlock: Simple human experiments regarding locally observed collective knowledge," tech. rep., US Army Research Laboratory Aberdeen Proving Ground, United States, 2015.

[16] L. Bradeško, M. Witbrock, J. Starc, Z. Herga, M. Grobelnik, and D. Mladenić, "Curious cat–mobile, context-aware conversational crowdsourcing knowledge acquisition," *ACM Transactions on Information Systems (TOIS)*, vol. 35, no. 4, pp. 1–46, 2017.

[17] S. C. Feng, W. Z. Bernstein, T. Hedberg, and A. Barnard Feeney, "Toward knowledge management for smart manufacturing," *Journal of computing and information science in engineering*, vol. 17, no. 3, 2017.

[18] T. R. Gruber, "A translation approach to portable ontology specifications," *Knowledge acquisition*, vol. 5, no. 2, pp. 199–220, 1993.

[19] M. Garetti, L. Fumagalli, and E. Negri, "Role of ontologies for cps implementation in manufacturing," *Management and Production Engineering Review*, 2015.

[20] B. R. Ferrer, W. M. Mohammed, M. Ahmad, S. Iarovyi, J. Zhang, R. Harrison, and J. L. M. Lastra, "Comparing ontologies and databases: a critical review of lifecycle engineering models in manufacturing," *Knowledge and Information Systems*, pp. 1–34, 2021.

[21] T. R. Gruber, "Toward principles for the design of ontologies used for knowledge sharing?," *International journal of human-computer studies*, vol. 43, no. 5-6, pp. 907–928, 1995.

[22] L. Lin, W. Zhang, Y. Lou, C. Chu, and M. Cai, "Developing manufacturing ontologies for knowledge reuse in distributed manufacturing environment," *International Journal of Production Research*, vol. 49, no. 2, pp. 343–359, 2011.

[23] S. An, P. Martinez, R. Ahmad, and M. Al-Hussein, "Ontology-based knowledge modeling for frame assemblies manufacturing," in *ISARC. Proceedings of the International Symposium on Automation and Robotics in Construction*, vol. 36, pp. 709–715, IAARC Publications, 2019.

[24] V. Zaletelj, E. Hozdić, P. Butala, *et al.*, "A foundational ontology for the modelling of manufacturing systems," *Advanced Engineering Informatics*, vol. 38, pp. 129–141, 2018.

[25] P. Grenon and B. Smith, "Snap and span: Towards dynamic spatial ontology," *Spatial cognition and computation*, vol. 4, no. 1, pp. 69–104, 2004.

[26] I. Niles and A. Pease, "Towards a standard upper ontology," in *Proceedings of the international conference on Formal Ontology in Information Systems-Volume 2001*, pp. 2–9, 2001.

[27] A. Gangemi, N. Guarino, C. Masolo, A. Oltramari, and L. Schneider, "Sweetening ontologies with dolce," in *International Conference on Knowledge Engineering and Knowledge Management*, pp. 166–181, Springer, 2002.

[28] H. Herre, "General formal ontology (gfo): A foundational ontology for conceptual modelling," in *Theory and applications of ontology: computer applications*, pp. 297–345, Springer, 2010.

[29] L. Schneider, "Designing foundational ontologies," in *International Conference on Conceptual Modeling*, pp. 91–104, Springer, 2003.

[30] M. Garetti and L. Fumagalli, "P-pso ontology for manufacturing systems," *IFAC Proceedings Volumes*, vol. 45, no. 6, pp. 449–456, 2012.

[31] Z. Usman, R. I. Young, N. Chungoora, C. Palmer, K. Case, and J. A. Harding, "Towards a formal manufacturing reference ontology," *International Journal of Production Research*, vol. 51, no. 22, pp. 6553–6572, 2013.

[32] E. Negri, L. Fumagalli, M. Macchi, and M. Garetti, "Ontology for service-based control of production systems," in *IFIP International Conference on Advances in Production Management Systems*, pp. 484–492, Springer, 2015.

[33] H.-K. Lin and J. A. Harding, "A manufacturing system engineering ontology model on the semantic web for inter-enterprise collaboration," *Computers in Industry*, vol. 58, no. 5, pp. 428–437, 2007.

[34] L. Temal, A. Rosier, O. Dameron, and A. Burgun, "Mapping bfo and dolce," in *MedInfo*, pp. 1065–1069, 2010.

[35] M. Vegetti, G. P. Henning, and H. P. Leone, "Product ontology: definition of an ontology for the complex product modelling domain," in *Proceedings of the Mercosur Congress on Process Systems Engineering*, 2005.

[36] H. Panetto, M. Dassisti, and A. Tursi, "Onto-pdm: Product-driven ontology for product data management interoperability within manufacturing process environment," *Advanced Engineering Informatics*, vol. 26, no. 2, pp. 334–348, 2012.

[37] C. Schlenoff, F. Tissot, J. Valois, and J. Lee, *The process specification language (PSL) overview and version 1.0 specification*. Citeseer, 2000.

[38] E. Järvenpää, N. Siltala, O. Hylli, and M. Lanz, "The development of an ontology for describing the capabilities of manufacturing resources," *Journal of Intelligent Manufacturing*, vol. 30, no. 2, pp. 959–978, 2019.

[39] S. Jaskó, A. Skrop, T. Holczinger, T. Chován, and J. Abonyi, "Development of manufacturing execution systems in accordance with industry 4.0 requirements: A review of standard-and ontology-based methodologies and tools," *Computers in Industry*, vol. 123, p. 103300, 2020.

[40] S. Borgo and P. Leitão, "Foundations for a core ontology of manufacturing," in *Ontologies*, pp. 751–775, Springer, 2007.

[41] P. Leitão, A. W. Colombo, and F. J. Restivo, "Adacor: A collaborative production automation and control architecture," *IEEE Intelligent Systems*, vol. 20, no. 1, pp. 58–66, 2005.

[42] D. Šormaz and A. Sarkar, "Simpm–upper-level ontology for manufacturing process plan network generation," *Robotics and Computer-Integrated Manufacturing*, vol. 55, pp. 183–198, 2019.

[43] Y. Lu, H. Panetto, Y. Ni, and X. Gu, "Ontology alignment for networked enterprise information system interoperability in supply chain environment," *International Journal of Computer Integrated Manufacturing*, vol. 26, no. 1–2, pp. 140–151, 2013.

[44] Y. Lu, H. Wang, and X. Xu, "Manuservice ontology: a product data model for service-oriented business interactions in a cloud manufacturing environment," *Journal of Intelligent Manufacturing*, vol. 30, no. 1, pp. 317–334, 2019.

[45] E. M. Sanfilippo, F. Belkadi, and A. Bernard, "Ontology-based knowledge representation for additive manufacturing," *Computers in Industry*, vol. 109, pp. 182–194, 2019.

[46] M. M. Ali, R. Rai, J. N. Otte, and B. Smith, "A product life cycle ontology for additive manufacturing," *Computers in Industry*, vol. 105, pp. 191–203, 2019.

[47] D. J. Russomanno, C. R. Kothari, and O. A. Thomas, "Building a sensor ontology: A practical approach leveraging iso and ogc models," in *IC-AI*, pp. 637–643, Citeseer, 2005.

[48] H. Neuhaus and M. Compton, "The semantic sensor network ontology," in *AGILE workshop on challenges in geospatial data harmonisation, Hannover, Germany*, pp. 1–33, 2009.

[49] Y. Hu, Z. Wu, and M. Guo, "Ontology driven adaptive data processing in wireless sensor networks," in *Proceedings of the 2nd international conference on Scalable information systems*, pp. 1–2, 2007.

[50] M. Compton, H. Neuhaus, K. Taylor, and K.-N. Tran, "Reasoning about sensors and compositions," *SSN*, vol. 522, pp. 33–48, 2009.

[51] M. Uschold and M. King, *Towards a methodology for building ontologies*. Citeseer, 1995.

[52] M. Uschold, "Building ontologies: Towards a unified methodology," in *Proceedings of 16th Annual Conference of the British Computer Society Specialists Group on Expert Systems*, Citeseer, 1996.

[53] M. Fernández-López, A. Gómez-Pérez, and N. Juristo, "Methontology: from ontological art towards ontological engineering," 1997.

[54] H. M. Kim, M. S. Fox, and M. Gruninger, "An ontology of quality for enterprise modelling," in *Proceedings 4th IEEE Workshop on Enabling Technologies: Infrastructure for Collaborative Enterprises (WET ICE'95)*, pp. 105–116, IEEE, 1995.

[55] M. Courtot, F. Gibson, A. L. Lister, J. Malone, D. Schober, R. R. Brinkman, and A. Ruttenberg, "Mireot: The minimum information to reference an external ontology term," *Applied Ontology*, vol. 6, no. 1, pp. 23–33, 2011.

[56] S. Ahmed, S. Kim, and K. M. Wallace, "A methodology for creating ontologies for engineering design," 2007.

[57] Z. Li, M. C. Yang, and K. Ramani, "A methodology for engineering ontology acquisition and validation," *Artificial Intelligence for Engineering Design, Analysis and Manufacturing: AI EDAM*, vol. 23, no. 1, p. 37, 2009.

[58] F. Ameri, C. Urbanovsky, and C. McArthur, "A systematic approach to developing ontologies for manufacturing service modeling," in *Proceedings of the workshop on ontology and semantic web for manufacturing*, vol. 14, 2012.

[59] D. Kiritsis, S. El Kadiri, A. Perdikakis, A. Milicic, D. Alexandrou, and K. Pardalis, "Design of fundamental ontology for manufacturing product lifecycle applications," in *IFIP International Conference on Advances in Production Management Systems*, pp. 376–382, Springer, 2012.

[60] M. C. Suárez-Figueroa, A. Gómez-Pérez, Ó. Muñoz-García, and M. Vigo, "gontt, a tool for scheduling and executing ontology development projects," in *SEKE*, pp. 614–619, 2010.

[61] X. Chang, R. Rai, and J. Terpenny, "Development and utilization of ontologies in design for manufacturing," *Journal of Mechanical Design*, vol. 132, no. 2, 2010.

[62] S. Akmal and R. Batres, "A methodology for developing manufacturing process ontologies," *Journal of Japan Industrial Management Association*, vol. 64, no. 2E, pp. 303–316, 2013.

[63] C. Hildebrandt, A. Köcher, C. Küstner, C.-M. López-Enríquez, A. W. Müller, B. Caesar, C. S. Gundlach, and A. Fay, "Ontology building for cyber–physical systems: Application in the manufacturing domain,"

IEEE Transactions on Automation Science and Engineering, vol. 17, no. 3, pp. 1266–1282, 2020.

[64] S. Ji, S. Pan, E. Cambria, P. Marttinen, and S. Y. Philip, "A survey on knowledge graphs: Representation, acquisition, and applications," *IEEE Transactions on Neural Networks and Learning Systems*, 2021.

[65] H. Paulheim, "Knowledge graph refinement: A survey of approaches and evaluation methods," *Semantic web*, vol. 8, no. 3, pp. 489–508, 2017.

[66] N. Noy, Y. Gao, A. Jain, A. Narayanan, A. Patterson, and J. Taylor, "Industry-scale knowledge graphs: Lessons and challenges: Five diverse technology companies show how it's done," *Queue*, vol. 17, no. 2, pp. 48–75, 2019.

[67] A. Hogan, E. Blomqvist, M. Cochez, C. d'Amato, G. de Melo, C. Gutierrez, J. E. L. Gayo, S. Kirrane, S. Neumaier, A. Polleres, *et al.*, "Knowledge graphs," *arXiv preprint arXiv:2003.02320*, 2020.

[68] Q. Wang, Z. Mao, B. Wang, and L. Guo, "Knowledge graph embedding: A survey of approaches and applications," *IEEE Transactions on Knowledge and Data Engineering*, vol. 29, no. 12, pp. 2724–2743, 2017.

[69] M. Färber, F. Bartscherer, C. Menne, and A. Rettinger, "Linked data quality of dbpedia, freebase, opencyc, wikidata, and yago," *Semantic Web*, vol. 9, no. 1, pp. 77–129, 2018.

[70] R. Y. Wang and D. M. Strong, "Beyond accuracy: What data quality means to data consumers," *Journal of management information systems*, vol. 12, no. 4, pp. 5–33, 1996.

[71] X. Dong, E. Gabrilovich, G. Heitz, W. Horn, N. Lao, K. Murphy, T. Strohmann, S. Sun, and W. Zhang, "Knowledge vault: A web-scale approach to probabilistic knowledge fusion," in *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 601–610, 2014.

[72] L. He and P. Jiang, "Manufacturing knowledge graph: a connectivism to answer production problems query with knowledge reuse," *IEEE Access*, vol. 7, pp. 101231–101244, 2019.

[73] H. Yan, J. Yang, and J. Wan, "Knowime: A system to construct a knowledge graph for intelligent manufacturing equipment," *IEEE Access*, vol. 8, pp. 41805–41813, 2020.

[74] C. Lv, Y. Lu, X. Yan, W. Lu, and H. Tan, "Supplier recommendation based on knowledge graph embedding," in *2020 Management Science Informatization and Economic Innovation Development Conference (MSIEID)*, pp. 514–518, IEEE, 2020.

[75] Y. Ding, W. Xu, Z. Liu, Z. Zhou, and D. T. Pham, "Robotic task oriented knowledge graph for human-robot collaboration in disassembly," *Procedia CIRP*, vol. 83, pp. 105–110, 2019.

[76] S. Munir, S. I. Jami, and S. Wasi, "Knowledge graph based semantic modeling for profiling in industry 4.0," *International Journal on Information Technologies & Security*, vol. 12, no. 1, 2020.

[77] A. Rivas, I. Grangel-González, D. Collarana, J. Lehmann, and M.-E. Vidal, "Unveiling relations in the industry 4.0 standards landscape based on knowledge graph embeddings," in *International Conference on Database and Expert Systems Applications*, pp. 179–194, Springer, 2020.

[78] I. Grangel-González, *A knowledge graph based integration approach for industry 4.0.* PhD thesis, Universitäts-und Landesbibliothek Bonn, 2019.

[79] P. Kumar and A. Gupta, "Active learning query strategies for classification, regression, and clustering: A survey," *Journal of Computer Science and Technology*, vol. 35, no. 4, pp. 913–945, 2020.

[80] C. Schröder and A. Niekler, "A survey of active learning for text classification using deep neural networks," *arXiv preprint arXiv:2008.07267*, 2020.

[81] S. Budd, E. C. Robinson, and B. Kainz, "A survey on active learning and human-in-the-loop deep learning for medical image analysis," *Medical Image Analysis*, p. 102062, 2021.

[82] V. Samsonov, J. Lipp, P. Noodt, A. F. Solvay, and T. Meisen, "More machine learning for less: Comparing data generation strategies in mechanical engineering and manufacturing," in *2019 IEEE Symposium Series on Computational Intelligence (SSCI)*, pp. 799–807, IEEE, 2019.

[83] B. Settles, "Active learning literature survey," 2009.

[84] D. Angluin, "Queries and concept learning," *Machine learning*, vol. 2, no. 4, pp. 319–342, 1988.

[85] R. Schumann and I. Rehbein, "Active learning via membership query synthesis for semi-supervised sentence classification," in *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pp. 472–481, 2019.

[86] D. A. Cohn, Z. Ghahramani, and M. I. Jordan, "Active learning with statistical models," *Journal of artificial intelligence research*, vol. 4, pp. 129–145, 1996.

[87] Y. Fu, X. Zhu, and B. Li, "A survey on instance selection for active learning," *Knowledge and information systems*, vol. 35, no. 2, pp. 249–283, 2013.

[88] J.-J. Zhu and J. Bento, "Generative adversarial active learning," *arXiv preprint arXiv:1702.07956*, 2017.

[89] D. Mahapatra, B. Bozorgtabar, J.-P. Thiran, and M. Reyes, "Efficient active learning for image classification and segmentation using a sample selection and conditional generative adversarial network," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 580–588, Springer, 2018.

[90] S. Sinha, S. Ebrahimi, and T. Darrell, "Variational adversarial active learning," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5972–5981, 2019.

[91] S. Ebrahimi, W. Gan, D. Chen, G. Biamby, K. Salahi, M. Laielli, S. Zhu, and T. Darrell, "Minimax active learning," *arXiv preprint arXiv:2012.10467*, 2020.

[92] C. Mayer and R. Timofte, "Adversarial sampling for active learning," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 3071–3079, 2020.

[93] N. Houlsby, F. Huszár, Z. Ghahramani, and M. Lengyel, "Bayesian active learning for classification and preference learning," *arXiv preprint arXiv:1112.5745*, 2011.

[94] T. Tran, T.-T. Do, I. Reid, and G. Carneiro, "Bayesian generative active deep learning," in *International Conference on Machine Learning*, pp. 6295–6304, PMLR, 2019.

[95] B. Boecking, W. Neiswanger, E. Xing, and A. Dubrawski, "Interactive weak supervision: Learning useful heuristics for data labeling," *arXiv preprint arXiv:2012.06046*, 2020.

[96] A. Awasthi, S. Ghosh, R. Goyal, and S. Sarawagi, "Learning from rules generalizing labeled exemplars," *arXiv preprint arXiv:2004.06025*, 2020.

[97] S. Biegel, R. El-Khatib, L. O. V. B. Oliveira, M. Baak, and N. Aben, "Active weasul: Improving weak supervision with active learning," *arXiv preprint arXiv:2104.14847*, 2021.

[98] L. Meng, B. McWilliams, W. Jarosinski, H.-Y. Park, Y.-G. Jung, J. Lee, and J. Zhang, "Machine learning in additive manufacturing: A review," *Jom*, vol. 72, no. 6, pp. 2363–2377, 2020.

[99] K. van Garderen, "Active learning for overlay prediction in semi-conductor manufacturing," 2018.

[100] W. Dai, A. Mujeeb, M. Erdt, and A. Sourin, "Towards automatic optical inspection of soldering defects," in *2018 International Conference on Cyberworlds (CW)*, pp. 375–382, IEEE, 2018.

[101] X. Yue, Y. Wen, J. H. Hunt, and J. Shi, "Active learning for gaussian process considering uncertainties with application to shape control of composite fuselage," *IEEE Transactions on Automation Science and Engineering*, 2020.

[102] Z. Lv, L. Wang, Z. Han, J. Zhao, and W. Wang, "Surrogate-assisted particle swarm optimization algorithm with pareto active learning for expensive multi-objective optimization," *IEEE/CAA Journal of Automatica Sinica*, vol. 6, no. 3, pp. 838–849, 2019.

[103] C. Pradel, D. Sileo, Á. Rodrigo, A. Peñas, and E. Agirre, "Question answering when knowledge bases are incomplete," in *International Conference of*

the Cross-Language Evaluation Forum for European Languages, pp. 43–54, Springer, 2020.

[104] A. D. Preece, "A new approach to detecting missing knowledge in expert system rule bases," *International journal of man-machine studies*, vol. 38, no. 4, pp. 661–688, 1993.

[105] J. Xu, W. Zhang, A. Alawini, and V. Tannen, "Provenance analysis for missing answers and integrity repairs," *IEEE Data Eng. Bull.*, vol. 41, no. 1, pp. 39–50, 2018.

[106] M. J. Franklin, D. Kossmann, T. Kraska, S. Ramesh, and R. Xin, "Crowddb: answering queries with crowdsourcing," in *Proceedings of the 2011 ACM SIGMOD International Conference on Management of data*, pp. 61–72, 2011.

[107] U. Waltinger, D. Tecuci, M. Olteanu, V. Mocanu, and S. Sullivan, "Usi answers: Natural language question answering over (semi-) structured industry data," in *Proceedings of the Twenty-Seventh AAAI Conference on Artificial Intelligence*, pp. 1471–1478, 2013.

[108] A. Bordes, J. Weston, and N. Usunier, "Open question answering with weakly supervised embedding models," in *Joint European conference on machine learning and knowledge discovery in databases*, pp. 165–180, Springer, 2014.

[109] J. M. Rožanec, P. Zajec, K. Kenda, I. Novalija, B. Fortuna, and D. Mladenić, "Xai-kg: knowledge graph to support xai and decision-making in manufacturing," 2021.

[110] H. Kagermann, "Change through digitization—value creation in the age of industry 4.0," in *Management of permanent change*, pp. 23–45, Springer, 2015.

[111] P. Zajec, J. M. Rožanec, I. Novalija, B. Fortuna, D. Mladenić, and K. Kenda, "Towards active learning based smart assistant for manufacturing," *arXiv preprint arXiv:2103.16177*, 2021.

[112] M. Alkahtani, A. Choudhary, A. De, and J. A. Harding, "A decision support system based on ontology and data mining to improve design using warranty data," *Computers & industrial engineering*, vol. 128, pp. 1027–1039, 2019.

[113] A. Giovannini, A. Aubry, H. Panetto, M. Dassisti, and H. El Haouzi, "Ontology-based system for supporting manufacturing sustainability," *Annual Reviews in Control*, vol. 36, no. 2, pp. 309–317, 2012.

[114] Q. Cao, F. Giustozzi, C. Zanni-Merk, F. de Bertrand de Beuvron, and C. Reich, "Smart condition monitoring for industry 4.0 manufacturing processes: An ontology-based approach," *Cybernetics and Systems*, vol. 50, no. 2, pp. 82–96, 2019.

[115] M. Elahi, F. Ricci, and N. Rubens, "A survey of active learning in collaborative filtering recommender systems," *Computer Science Review*, vol. 20, pp. 29–50, 2016.

[116] S. C.-H. Yang, C. Rank, J. Whritner, O. Nasraoui, and P. Shafto, "Unifying recommendation and active learning for information filtering and recommender systems," 2020.

[117] D. Pelleg and A. Moore, "Active learning for anomaly and rare-category detection," *Advances in neural information processing systems*, vol. 17, pp. 1073–1080, 2004.

[118] N. Abe, B. Zadrozny, and J. Langford, "Outlier detection by active learning," in *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 504–509, 2006.

[119] Y. Liu, Z. Li, C. Zhou, Y. Jiang, J. Sun, M. Wang, and X. He, "Generative adversarial active learning for unsupervised outlier detection," *IEEE Transactions on Knowledge and Data Engineering*, vol. 32, no. 8, pp. 1517–1528, 2019.

DOI: 10.1561/9781680838770.ch4

Chapter 4

Multimodal Human Machine Interactions in Industrial Environments

By Rubén Alonso, Nino Cauli and Diego Reforgiato Recupero

This chapter will present a review of Human Machine Interaction techniques for industrial applications. A set of recent HMI techniques will be provided with emphasis on multimodal interaction with industrial machines and robots. This list will include Natural Language Processing techniques and others that make use of various complementary interfaces: audio, visual, haptic or gestural, to achieve a more natural human-machine interaction. This chapter will also focus on providing examples and use cases in fields related to multimodal interaction in manufacturing, such as augmented reality. Accordingly, the chapter will present the use of Artificial Intelligence and Multimodal Human Machine Interaction in the context of STAR applications.

4.1 Introduction

Since the beginning of the 20th century, automation played a fundamental role in the manufacturing industry (Wang, 2019). Starting from the sixties, robots were introduced in factories speeding up the manufacturing process. Initially there were strict boundaries between robots' and humans' work-spaces. In order to avoid injuries, workers were not allowed to enter in the robots' working space. Unfortunately, this rigid organization has its limitations. Both robots and humans excel in different areas and a proper collaboration between them can result in a more efficient assembling process. Robots are faster, stronger and more precise in repetitive assembling tasks, while humans are better in decision making and they can easily adapt to unexpected situations. The exponential improvements achieved in the 21st century in AI, perception algorithms and robot control, gradually allowed for a shared work-space between human workers and robots.

Robots use on-board and external sensors to be aware of their surrounding environment. The data output of sensors range from simple single dimension data (contact sensors, ultrasonic distance sensors) to complex high dimensions data (microphones, lidar sensors, RGB cameras, depth cameras). In order to have a better interaction with human workers and other machines in the factory, robots need to merge the information received by every kind of sensor available. This multimodal interaction exists in both ways: while interacting with robots, human workers must not be limited to a restricted group of modalities and devices (keyboards, mouse, screen), but they should be able to use all the modalities made available by their bodies (speech, vision, gestures, touch).

The goal of this chapter is to present the various types of multimodal interaction in industrial environments. After introducing the problem of multimodal interaction we will present some examples of modalities for a natural interaction between human workers and robots/machines such as speech (intended also as Natural Language Processing of text obtained using speech-to-text tools) and vision. We will then make a step further to the idea of multimodal interaction introducing the concept of Extended Reality (XR), where a human is able to remotely control a robot sharing its sensory stimuli.

More specifically, the remainder of this chapter is organized as it follows. Section 4.2 includes all the possible kinds of multimodal interaction between humans and machines. Section 4.3 describes how NLP techniques can be employed within the manufacturing domain. Section 4.4 illustrates human motion recognition and prediction for human robot interaction in manufacturing industry. Section 4.5 illustrates XR technologies which include augmented, mixed and virtual reality. Moreover, a use case showing the application of virtual reality to remote-control a humanoid robot within the manufacturing domain is presented as well. Finally, Section 4.6 concludes the paper.

4.2 Multimodal Interaction

Since the presentation of the famous Put-That-There (Bolt, 1980), innumerable papers have been written about the advantages and disadvantages, problems and solutions aroused from the natural interaction between humans and machines. Multimodal Interaction discipline is based on the idea that human communication is multimodal. Thus, if hoping to interact with machines in the same way as it is done with humans, the interaction must not be limited to a group of modalities and devices, as it has been done until now, using mainly keyboard and mouse as data input and graphical representations as data output.

Some authors (Waibel *et al.*, 1996) point out that it is not advisable to reduce the interaction exclusively to human ↔ machine. They classify the multimodal interaction interfaces in five different classes:

- Human → Machine: in an unidirectional way, as data input mode. For example, a user dictating a text to the computer or giving orders to a robot (without receiving any complex feedback).
- Human ↔ Machine: in a bidirectional and interactive way between the human and the machine, like, for example, in a route planner.
- Human ↔ Multimedia Data: as the extraction of data from multimedia information. For example, the extraction of meaningful images and the transcription of text from video-recorded news, for the subsequent search by a human.
- Human ↔ Machine ↔ Human: where the machine mediates in the interaction between two humans that do not have the same knowledge, lack part of the context or simply because they are far from each other and cannot interact directly.
- Human ↔ Human (observed and assisted by machine): it is not mediated by a machine, but there exists one for assisting the user. For example, a system that records and transcribes meetings which can be searched later looking for actions defined in previous meetings.

In (Alonso and Torres, 2010) the authors extended the list to support a new category: **Human ↔ Multiple Machines**, where the user interacts in a multimodal way with a group of programmable machines, such as robots, using different media and devices, and collaborates with all of them.

This theoretical classification, essential to understand multimodal interaction, is somewhat diluted in practice, especially after the emergence of XR technologies. Anyway, the six classes are relevant to the manufacturing industry, and all have been addressed in a certain degree in the literature related to Artificial Intelligence (AI) and Human Machine Interaction (HMI) in recent years.

For example, Roitberg *et al.* (Roitberg *et al.*, 2015) present an interesting approach for improving the efficiency of Human-Robot interaction. This approach is based on multimodal interfaces, and is focused on the industrial environment. Their research is based on monitoring and interpreting human operations, using video depth information provided by different sensors. They use Microsoft Kinect v2 for skeleton tracking, Asus Xtion PRO for object tracking and Leap Motion for hand and finger pose tracking.

Liu *et al.* (Liu *et al.*, 2018) focus on multimodal human ↔ robot collaboration, especially in repetitive and dangerous tasks. They suggest that the more modalities are included and fused, the more robust the collaboration will be. For this purpose, they present an architecture and a use case for operator-robot collaboration in which body motion recognition, hand motion recognition and speech commands recognition are combined.

Concerning the use of multimodal interaction for operator training, (Vélaz *et al.*, 2014) analysed the influence of four interaction technologies and modalities (including mouse, haptic systems and 2D and 3D position capture) for the learning of a procedural assembly task. Among its conclusions it is worth noting that the results showed that the differences between the training performed with these interaction technologies were not significantly different from the traditional training performed by the operators.

Another significant example of multimodal interaction with multiple machines that could be extrapolated to the manufacturing sector is the coordination of multiple unmanned aerial vehicles. Several authors (e.g.: Cacace *et al.*, 2016b, Cacace *et al.*, 2016a) are working on the coordination of machines, using the information obtained through different modalities to solve interaction and coordination problems.

The improvement of recognition thanks to multimodal interaction has been proven in many studies (e.g.: Kettebekov *et al.*, 2002, Oviatt *et al.*, 2003) where the benefits of multimodal HMI were demonstrated for completing the available information and improve the recognition ratio using supporting modalities.

4.3 Employment of Natural Language Processing Within Manufacturing

Natural Language Processing (NLP) is a subset of AI that helps identifying key elements from human instructions, extract relevant information and process them in a manner that machines can understand. Integrating NLP technologies into the system helps machines understand human language and mimic human behaviour. For example, Amazon's Echo, Microsoft's Cortana and Apple's Siri make an extensive use of NLP technologies to interact with the users.

NLP technologies speed up the operation of a whole system cutting down the response time. Imagine a scenario where a manufacturing company hires a data scientist to collect and analyse all the machine readings, reporting any sort of problems. One disadvantage to this scheme is that by the time the management reads the report one problem might have happened causing damage to the entire process. If a robot with sensors and NLP technologies embedded is employed, this might remotely access the machines and detect in real time any change or problem providing an action to be executed. The robot might even communicate with users and accept input in natural language. Therefore, by leveraging NLP technologies, the middleman can be cut out while at the same time keeping the system effective.

Within the manufacturing industry the NLP might be adopted for the following tasks:

- **Process Automation:** The use of NLP technologies in the manufacturing process allows the automatic execution of repetitive tasks like paperwork and report analysis (e.g., Cristian *et al.*, 2019). Besides, it benefits the workflow of the entire process as each employee can be focused on tasks which require human intervention and capabilities. Authors in (Kang *et al.*, 2019) developed the feedback generation method based on Constraint-based Modeling (CBM) coupled with NLP and domain ontology, designed to support formal manufacturing rule extraction. In detail, the developed method identifies the necessity of input text validation based on the predefined constraints and provides the relevant feedback to help the user modify the input text, so that the desired rule can be extracted.
- **Inventory Management[1]:** Analysing data about the sales of certain products is essential to assess the correct decisions for a company to optimize and maximize profits. By leveraging NLP technologies the resulting benefits are: (1) the entire process becomes more comprehensive; (2) it is more difficult to incur errors related to the analysis of sales; (3) it is easier to analyse the manufactured products and discard those with low quality without affecting the supply chain and sales. On a different level, authors in (Vicari and Gaspari, 2020, Carta *et al.*, 2021) have employed NLP and Machine Learning techniques to automatically identify patterns, sentiment or other elements within a text which might be correlated to the stock variation.

[1] https://cmr.berkeley.edu/2021/01/managing-supply-chain-risk/

- **Emotional Mapping:** Sentiment analysis and emotion detection (Atzeni *et al.*, 2018, Atzeni and Recupero, 2020) are one of the most exciting features of NLP. Early NLP systems allowed organizations to collect speech-to-text communication without accurately determining its full meaning. Today, NLP approaches can sort and understand the nuances and emotions in human voices and text, giving organizations unparalleled insight. Learning customer expectations is a very important element in manufacturing. NLP technologies permit to identify emotions and opinions of customers (Dridi *et al.*, 2019, Recupero *et al.*, 2015) and provide actions to improve products and the selling process. Knowing the expectations of customers is key to build a longer relationship and create engagement with them.
- **Operation Optimization:** Furthermore, NLP technologies can be employed to trace the performance of equipment, identifying potential inefficiency. This enables a detailed monitoring of the machinery and taking measures to improve the overall system operability. A review of machine learning approaches for the optimization of production processes covers the majority of relevant literature from 2008 to 2018 dealing with machine learning and optimization approaches for product quality or process improvement in the manufacturing industry (Weichert *et al.*, 2019).

4.4 Human Motion Recognition and Prediction for Human Robot Interaction in Manufacturing

In order to safely interact with humans, robots need to understand human intentions and predict their movements. With the ability to recognise and to predict human actions, industrial robots are able to avoid dangerous collisions and to improve collaborative work anticipating some actions (i.e. passing to the worker the proper tool based on the predicted worker's action).

4.4.1 Video Action Recognition and Prediction

Human action recognition is a complex task that needs as much information as possible about the subject performing the action. RGB and depth cameras are the most suitable sensors for this task: a video sequence of a human performing an action carries information about his visual appearance, the context of the action and the motion of his body.

In order to recognise human actions from images, two steps are needed: action representation and action classification (Kong and Fu, 2018). Traditionally, handcrafted features are used to represent the actions (Jia and Yeung, 2008, Yuan *et al.*, 2016), and standard classifiers are used to recognise the action (e.g. SVN, k-means). The representation of the actions can vary from low level features (edges, corners) to high level ones (body shape, skeletal information). Choosing the optimal handcrafted features that best suit the task of action recognition can be tricky. Automatically extracted features are often more robust and achieve better performances. The recent increase in computational power brought to the rise of Convolutional Neural Networks (CNNs). CNNs are a type of Deep Artificial Neural Networks (DNNs) where for each of the several layers is applied a convolution between 2D weights kernels and the 2D channels of the previous layer. The output of each layer are 2D feature maps extracted from the previous layer (low level features for the initial layers and high level ones for the last layers). With their deep structure and with enough training data, CNNs are able to generate features for action recognition that outperform handcrafted ones. CNNs are frequently used to extract features to represent actions, achieving state-of-the-art results (Kong and Fu, 2018, Özyer *et al.*, 2021).

CNNs are data driven models and one of their drawbacks is the need of big labelled datasets with high quality images. The following are some examples of popular datasets for video action recognition, for a more exhaustive list please refer to (Kong and Fu, 2018, Özyer *et al.*, 2021):

- **UCF-101 (Soomro *et al.*, 2012):** One of the most used datasets for video action recognition. UCF-101 is a large dataset with 13,320 different YouTube videos from 101 categories. This dataset has high variability in camera angles, actors and backgrounds.
- **YouTube-8M (Abu-El-Haija *et al.*, 2016):** This is a very large multi-label video classification dataset (8 million videos for a total of 500K hours). The videos are extracted from YouTube and they are annotated with 4800 machine-generated labels.
- **The Kinetics Human Action Video Dataset (Kay *et al.*, 2017):** This dataset contains 306,245 YouTube clips of 10s each. The clips are grouped in 400 human action classes and are taken from different YouTube videos.
- **Moments in Time (Monfort *et al.*, 2019):** A large-scale human annotated dataset with one million videos of 3 seconds corresponding to dynamic events. Each video is labeled with one among 339 different classes.

While it is possible to recognize action from static images, they lack information about the motion during time. CNNs need to be extended in order to use the time information of video sequences. The most common approaches are the followings:

- **3D CNNs:** These networks are a particular type of CNNs composed by multiple layers of 3D convolutions obtained using 3D kernels. Receiving as input a sequence of frames stacked in one dimension, 3D CNNs are able to extract features related both to space and time. S. Ji *et al.* (Ji *et al.*, 2012) used 3D CNNs to recognize human actions in the real-world environment of airport surveillance videos. The authors compared their model with the state-of-the-art algorithms at the time achieving superior performance.

- **Multi-stream networks:** This type of architecture classifies its input merging together the output of several CNNs. Each CNN receives a different type of input. K. Simonyan and A. Zisserman (Simonyan and Zisserman, 2014) proposed a two-stream CNN for action recognition. The first stream received as input a single RGB frame, while the second stream received as input the multi-frame optical flow, carrying temporal information of the action. The authors tested the network on the UCF-101 dataset obtaining state-of-the-art results.
- **Recurrent neural networks (RNNs):** RNNs are special artificial neural network with internal loops in the connection between layers. Their special structure makes them able to keep a memory of the past and to generate an output based on the sequence of the most recent inputs received. J. Yue-Hei Ng *et al.* (Yue-Hei Ng *et al.*, 2015) introduced an hybrid network that joins together CNNs with RNNs. Their model is composed by GoogLeNet convolutional layers followed by 5 LSTM layers. In the paper the authors perform several ablation studies on a video recognition task showing advantages and disadvantages of using recurrent layers.

Video action recognition is the problem of recognising the action performed by a subject based on a video sequence of the entire movement. The problem of predicting the action performed based only on a video of an initial portion of the action is called action prediction. The most recent action/motion prediction systems tend to use the combination of CNNs and RNNs (Lee *et al.*, 2017), better suited for the analysis of video sequences. In Human Robot Collaboration (HRC) scenarios, the prediction of the type of action performed by the human might not be enough. Often the robot needs to know the full body motion during the next action performed by the human in order to successfully perform the collaborative task. Recently some researchers were able to predict the next frames of a motion based on the action to be performed and past frames (Finn *et al.*, 2016, Jung *et al.*, 2019).

For a Robot interacting with a dynamic environment, it is of primary importance being able to model the surroundings and to predict how the environment evolves through time. With a faithful representation of the environment, the robot is able to detect unexpected behaviours and to correct its actions accordingly. This idea is borrowed from cognitive science: in the Predictive Coding (Rao and Ballard, 1999) cognition theory, the brain is constantly predicting the sensory outcome (top-down process) and comparing it with the actual one. At the same time the error between predicted and actual sensory stimuli is back-propagated to the highest layers (bottom-up process) in order to revise and update the internal predictive models (a similar idea applied to robot control was studied under the name of Expected Perception (Barrera and Laschi, 2010, Cauli *et al.*, 2016)). Jun Tani implemented on robotics platforms several models based on the Predictive Coding paradigm (Tani, 2016). One of the most recent is the Predictive Visuo-Motor Deep Dynamic Neural Network (P-VMDNN) (Hwang *et al.*, 2018). This Deep-RNN model can be used both to predict the next RGB frames and encoders values during a motion, and to recognise an action performed by a human placed in front of the robot.

4.4.2 Video Action Recognition and Prediction for HRC in Manufacturing

In recent years we are seeing a gradual introduction of shared spaces and collaborative tasks between humans and robots in factories. Human and robotic workers can collaborate during the assembly process of specific components. In these scenarios, the robot must predict the human coworker action in order to plan its own motion. The application of video action recognition models to HRC in manufacturing is still a relatively new topic (Wang, 2019).

The most straightforward approaches use handcrafted features to represent the actions. E. Coupeté *et al.* (Coupeté *et al.*, 2019) extract the skeletal representation of the upper-torso of a worker from depth images. The sequence of skeletal position during a motion is given as input to an Hidden Markov Model in order to recognise the performed gesture. The model is tested in an assembly scenario where a worker and a robot collaborate to mount a mechanical piece.

A different approach is to automatically extract the best features using a CNN. P. Wang *et al.* (Wang *et al.*, 2018) use AlexNet to recognise specific gestures from a video of a worker assembling an engine. The convolutional layers extract the features while 3 fully connected layers classify the gesture. The architecture based the classification only on single frames.

We already mentioned that single images lack information of the temporal evolution of the action. Using both RGB images and optical flow as inputs solves the problem. Q. Xiong *et al.* (Xiong *et al.*, 2020) use the two-streams network proposed by (Simonyan and Zisserman, 2014) to recognise the actions from closeup videos of workers assembling engines' parts. The network has 2 CNN branches, one receiving as input RGB images and the other optical flow images. Due to the small size of the engine block assembly dataset used in the experiment, the authors apply transfer learning. They first pretrain the entire network on a bigger generic action dataset and then they finetune the last layers on the engine block assembly dataset.

RNNs are other models able to keep temporal information of the recently seen frames. Z. Liu *et al.* (Liu *et al.*, 2019) developed a system able to predict the next action performed by a worker while assembling a computer. A robot passes the worker the proper tool based on the predicted action. The authors use a CNN as feature extractor followed by an LSTM layer and a fully connected layer to classify the next action. The input of the system are the images from a top-down camera mounted above the working table.

While a fair amount of work on action recognition in manufacturing already exists, the problem of human motion prediction in HRC needs to be studied in more details. A robot able to predict in each instant where the body of the human co-worker will be, can easily avoid collision, spot mistakes and make recovering actions.

It is clear that CNNs are the most reliable tool for features extraction from videos. CNNs need a big amount of data to learn properly and be able to generalise. Unfortunately, not many datasets for video action recognition in factory assembly scenario exist (Kong and Fu, 2018, Özyer *et al.*, 2021). New specific video datasets are difficult to generate and the labelling process is highly time consuming. Domain transfer and simulated datasets are a valid solution to the problem. M. Fabbri *et al.* (Fabbri *et al.*, 2018) generated a big dataset for Multi-People Tracking using the Grand Teft Auto V game engine. Generating a simulated dataset is faster than collecting a real one and labelling is automatic. An action recognition model trained on a simulated dataset with high variability and realism is able to transfer the knowledge learned in simulation to the real world.

4.5 XR in Manufacturing Industry

XR related technologies are facilitating multimodal interaction in Industry 4.0 and thus enabling tangible in-site visualisations and interactions with industrial assets (Simões *et al.*, 2018).

The term XR can be considered as an umbrella for the terms augmented (AR), mixed (MR) and virtual (VR) reality, which differ in how much real and virtual content they display and the level of interactivity. As detailed in Alizadehsalehi *et al.*, 2020 VR is characterised by high virtual content and low interactivity, while AR is characterised by high real content and higher interactivity. MR lies in the middle of both, including higher levels of virtual and real content, and high interactivity.

4.5.1 Related Work of XR in Industry

The use of XR in industry has been suggested since the early 90's, where for example Thomas and David, 1992 proposed the superimposition of certain information on real world objects. Since that point there are hundreds of examples of XR aided manufacturing, Bottani and Vignali present an exhaustive list of them in their article *"Augmented reality technology in the manufacturing industry: A review of the last decade"* (Bottani and Vignali, 2019).

In addition to the Boeing article (Thomas and David, 1992) already mentioned above, for example Karlsson *et al.* (Karlsson *et al.*, 2017) suggest an approach for the presentation of superimposed information, e.g. information on potential bottlenecks, that can help decision making in manufacturing.

Workforce training is another activity where the use of XR is increasing, especially after the rise of robotic systems and complex machines in shopfloors. For example. safety training is another area where multimodal interaction and XR are absolutely worthwhile. As detailed in Doolani *et al.*, 2020, these systems reduce the risks of harm that can be caused by machines as well as damage to them, and offer a platform for learning-by-doing approach that can be used multiple times without worrying about the costs, availability or risks associated with the use of real machines.

The possibility of remote guidance is another advantage of XR systems in the manufacturing environment. For example Fast-Berglund *et al.*, 2018 validated a use case in which the expert uses AR to guide the novice operator in an assembly task and gives directions and corrections in case there is something wrong in the assembling. Their conclusion is that thanks to the AR being able to give instant feedback, it makes it practically impossible to do the assembly wrong and therefore the results are highly positive.

4.5.2 Use Case: Virtual Reality to Remote-control a Robot

In this section we are going to describe the work of authors in Alonso *et al.*, 2021 related to a general-purpose, open-source framework for teleoperating a NAO humanoid robot through a Virtual Reality (VR) headset. As the proposed architecture is general, it would be straightforward to replace the NAO robot with Kuka[2] or Universal Robot,[3] two well known robots used in several production environments around the world. The architecture presented in Alonso *et al.*, 2021 includes a VR

[2] http://www.kuka.com

[3] https://www.universal-robots.com

interface for the Oculus Rift[4] using the Unity game engine to perform robot actions through the VR controllers and exploits the flexibility of the Robot Operating System (ROS) for the control and synchronization of the robot hardware. This work gives ideas on potential architecture that can be employed within the manufacturing domain to allow the robots (e.g. Kuka or Universal Robot, both supported by ROS) to protect workers from repetitive, mundane, and dangerous tasks while also creating more desirable jobs such as engineering, programming, management and equipment maintenance. In the following we will show details of the tools used for their work. Let us first start giving some background information about the Unity, ROS and NAO software platforms.

Unity 3D[5] is a game engine which supports the development of 2D and 3D games, Virtual and Mixed Reality experiences and simulations.

ROS is an open-source framework for robot software whose architecture includes Nodes, Messages, Topics, Services, and Actions. Nodes are processes that carry out a computation. Messages are exchanged by nodes. A node sends a message by posting it on a certain topic. Services are needed by nodes that need to perform remote procedure calls. Actions are used to send a request to a node to perform a certain task for longer time and receive a reply. Then, ROS packages are a collection of code for easy reuse and stacks are a collection of packages that jointly offer some functionalities.

The authors employed NAO as the robotic platform but, as already mentioned, robots such as Kuka or Universal Robots may be employed. The Kuka system software is the operating software containing all the basic functions needed for the deployment of the robot system. Kuka robots come with a control panel with a display and axis control buttons and a 6D mouse which is used to manually move the robot. The control panel allows the users to view and create new and modify existing programs. A rugged computer lies in the control cabinet communicates with the robot system via the Multi Function Card, which controls the real-time servo drive electronics. Servo position feedback is transmitted to the controller through the DSE-Resolver Digital Converter/RDC connection. The software includes two elements running on parallel – the user interface and program storage. Figure 4.1 shows a Kuka robot palletizing food in a bakery. Universal robots consist of industrial collaborative robot arms (cobots), which are six-jointed robot arms with a very low weight (from 11 to 33 kilos) with a lifting ability from 3 to 16 kilos. These cobots can work right alongside personnel with no safety guarding, based on the

[4] https://www.oculus.com/rif t/

[5] https://unity.com/

Figure 4.1. A Kuka robot palletizing food in a bakery (taken from Wikipedia).

results of a mandatory risk assessment.[6] The robot arm can run in two operating modes of the safety functions; a normal and a reduced one. A switch between safety settings during the cobot's operation is also possible. Figure 4.2 shows a Universal Robot lifting an object.

In their work the authors show how through the remotes and the VR headset the VR interface allows the teleoperation of the NAO and the recording of a movements sequence for later execution. During the former, the user and the robot are not in the same room. Therefore, the user exploits the VR interface as a source of input and for having a visible and understandable representation of the remote robot status. The recording of a movements sequence allows the user to perform a number of tasks and save them in certain collections. Whenever needed, they can play them back.

As the ROS framework allows the development and run on different machines it is easier and more flexible to support both the storing and the playing of recorded actions of the robot.

Figure 4.3 illustrates the architecture of the VR system developed by the authors. It includes three main software components (VR, ROS and Rosbridge) and two hardware devices (Oculus Rift and NAO). The VR Component leverages the Unity

[6] https://www.iso.org/standard/62996.html

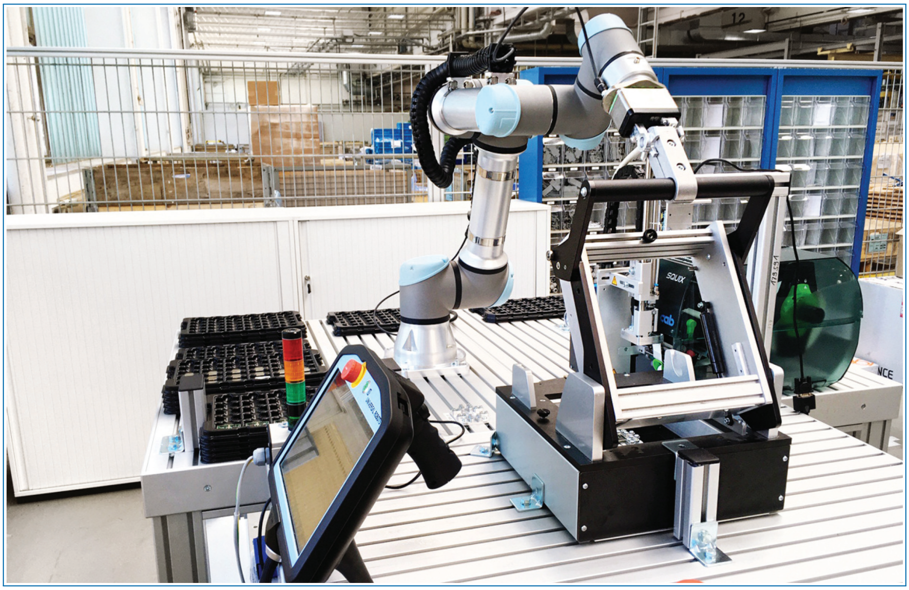


Figure 4.2. A Universal Robot lifting an object (taken from https://www.therobotreport.com/voith-robotics-cuts-ties-franka-emika-adds-universal-robots/).

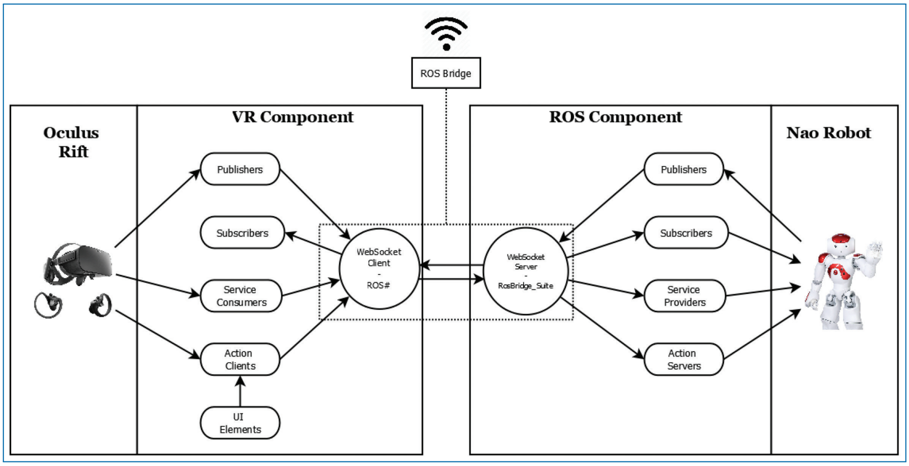


Figure 4.3. Architecture of the Virtual Reality system. Taken from Alonso *et al.*, 2021.

game engine for displaying the interface on the Oculus Rift. Unity has been chosen for the existing Oculus SDK that facilitates the developing process. The ROS component controls the robot through the VR simulation or the management of real hardware. It includes the ROS framework, multiple packages provided by ROS Nao Drivers, custom Publisher, Subscriber, Action Servers and Service Provider that have been implemented for supporting the VR control. The Ros Bridge is the connection between the VR and ROS components. It provides the methods for passing messages between them, for managing the information serialization and deserialization, and the connection and the delivery through WebSockets.

4.6 Conclusions

In this chapter we have presented various types of multimodal interaction within the manufacturing domain. First we have introduced the classification of multimodal interaction interfaces, indicating all the possible ways a user can interact with one or multiple machines. Then we briefly described the NLP research area and how it can be employed to automatically let an independent system (e.g., an agent or robot) to identify relevant information within the manufacturing. Next, we examined the ability of robots of recognising and predicting human actions by using cameras as sensors and deep learning as breakthrough machine learning technology. We continued discussing the XR related technologies (e.g., augmented, mixed, virtual reality) and how they can facilitate multimodal interaction in Industry 4.0. Finally, we showed an architecture of a use case where virtual reality technology has been adopted to remote-control a robot and how this schema can be adapted to be employed within the manufacturing domain.

Secure, safe, reliable AI systems in manufacturing environments, such as those investigated in the STAR project, can benefit from all of these technologies in their goal to make systems more trusted and human-centric. As part of the STAR project, research will continue on Human Robot Interaction and on knowledge systems, that benefit from NLP techniques and are accessible through multimodal interaction.

Acknowledgement

This work has been carried out in the H2020 STAR project, which has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 956573.

References

Abu-El-Haija, S., N. Kothari, J. Lee, P. Natsev, G. Toderici, B. Varadarajan, and S. Vijayanarasimhan (2016). "Youtube-8m: A large-scale video classification benchmark". *arXiv preprint arXiv:1609.08675*.

Alizadehsalehi, S., A. Hadavi, and J. C. Huang (2020). "From BIM to extended reality in AEC industry". *Automation in Construction*. 116: 103254.

Alonso, R. and M. I. Torres (2010). “Architecture for the Multimodal coordination of semi-autonomous agents”. In: *International Conference on Agents and Artificial Intelligence, ICAA 2010, Valencia, Spain. 22–24 January, 2010*.

Alonso, R., A. Bonini, D. Reforgiato Recupero, and D. Spano (2021). “Exploiting Virtual Reality and the Robot Operating Systemto remote-control a Humanoid Robot”. *Submitted to Multimedia Tools and Applications*.

Atzeni, M., A. Dridi, and D. R. Recupero (2018). “Using frame-based resources for sentiment analysis within the financial domain”. *Prog. Artif. Intell.* 7(4): 273–294. DOI: 10.1007/s13748- 018-0162-8. URL: https://doi.org/10.1007/s13748-018-0162-8.

Atzeni, M. and D. R. Recupero (2020). “Multi-domain sentiment analysis with mimicked and polarized word embeddings for human-robot interaction”. *Future Gener. Comput. Syst.* 110: 984–999. DOI: 10.1016/j.future.2019.10.012. URL: https://doi.org/10.1016/j.future.2019.10.012.

Barrera, A. and C. Laschi (2010). “Anticipatory visual perception as a bio-inspired mechanism underlying robot locomotion”. In: *2010 Annual International Conference of the IEEE Engineering in Medicine and Biology*. IEEE. 3206–3209.

Bolt, R. A. (1980). ““Put-that-there” Voice and gesture at the graphics interface”. In: *Proceedings of the 7th annual conference on Computer graphics and interactive techniques*. 262–270.

Bottani, E. and G. Vignali (2019). “Augmented reality technology in the manufacturing industry: A review of the last decade”. *IISE Transactions*. 51(3): 284–310.

Cacace, J., A. Finzi, and V. Lippiello (2016a). “Multimodal interaction with multiple co-located drones in search and rescue missions”. *arXiv preprint arXiv:1605.07316*.

Cacace, J., A. Finzi, V. Lippiello, M. Furci, N. Mimmo, and L. Marconi (2016b). “A control architecture for multiple drones operated via multimodal interaction in search & rescue mission”. In: *2016 IEEE International Symposium on Safety, Security, and Rescue Robotics (SSRR)*. IEEE. 233–239.

Carta, S., S. Consoli, L. Piras, A. S. Podda, and D. R. Recupero (2021). “Event detection in finance using hierarchical clustering algorithms on news and tweets”. *PeerJ Comput. Sci.* 7: e438. DOI: 10.7717/peerj-cs.438. URL: https://doi.org/10.7717/peerj-cs.438.

Cauli, N., E. Falotico, A. Bernardino, J. Santos-Victor, and C. Laschi (2016). “Correcting for changes: expected perception-based control for reaching a moving target”. *IEEE Robotics & Automation Magazine*. 23(1): 63–70.

Coupeté, E., F. Moutarde, and S. Manitsaris (2019). “Multi-users online recognition of technical gestures for natural human–robot collaboration in manufacturing”. *Autonomous Robots*. 43(6): 1309–1325.

Cristian, M., S. Christian, and T. Tudor (2019). “A Study in the Automation of Service Ticket Recognition using Natural Language Processing”. In: 1–6. DOI: 10.23919/SOFTCOM.2019.8903676.

Doolani, S., C. Wessels, V. Kanal, C. Sevastopoulos, A. Jaiswal, H. Nambiappan, and F. Makedon (2020). “A Review of Extended Reality (XR) Technologies for Manufacturing Training”. *Technologies*. 8(4): 77.

Dridi, A., M. Atzeni, and D. Reforgiato Recupero (2019). “FineNews: fine-grained semantic sentiment analysis on financial microblogs and news”. *International Journal of Machine Learning and Cybernetics*. 10(8): 2199–2207. DOI: 10.1007/s13042-018-0805-x. URL: https://doi.org/10.1007/s13042-018-0805-x.

Fabbri, M., F. Lanzi, S. Calderara, A. Palazzi, R. Vezzani, and R. Cucchiara (2018). “Learning to detect and track visible and occluded body joints in a virtual world”. In: *Proceedings of the European Conference on Computer Vision (ECCV)*. 430–446.

Fast-Berglund, Å., L. Gong, and D. Li (2018). “Testing and validating Extended Reality (xR) technologies in manufacturing”. *Procedia Manufacturing*. 25: 31–38.

Finn, C., I. Goodfellow, and S. Levine (2016). “Unsupervised learning for physical interaction through video prediction”. In: *Advances in neural information processing systems*. 64–72.

Hwang, J., J. Kim, A. Ahmadi, M. Choi, and J. Tani (2018). “Dealing with large-scale spatio-temporal patterns in imitative interaction between a robot and a human by using the predictive coding framework”. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*.

Ji, S., W. Xu, M. Yang, and K. Yu (2012). “3D convolutional neural networks for human action recognition”. *IEEE transactions on pattern analysis and machine intelligence*. 35(1): 221–231.

Jia, K. and D.-Y. Yeung (2008). “Human action recognition using local spatio-temporal discriminant embedding”. In: *2008 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE. 1–8.

Jung, M., T. Matsumoto, and J. Tani (2019). “Goal-Directed Behavior under Variational Predictive Coding: Dynamic Organization of Visual Attention and Working Memory”. *arXiv preprint arXiv:1903.04932*.

Kang, S., L. Patil, A. Rangarajan, A. Moitra, T. Jia, D. Robinson, and D. Dutta (2019). “Automated feedback generation for formal manufacturing rule extraction”. *Artificial Intelligence for Engineering Design, Analysis and Manufacturing*. 33(3): 289–301. DOI: 10.1017/S0890060419000027.

Karlsson, I., J. Bernedixen, A. H. Ng, and L. Pehrsson (2017). “Combining augmented reality and simulation-based optimization for decision support in manufacturing". In: *2017 Winter Simulation Conference (WSC)*. IEEE. 3988–3999.

Kay, W., J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev, *et al.* (2017). "The kinetics human action video dataset". *arXiv preprint arXiv:1705.06950*.

Kettebekov, S., M. Yeasin, N. Krahnstoever, and R. Sharma (2002). "Prosody based co-analysis of deictic gestures and speech in weather narration broadcast". In: *Workshop on Multimodal Resources and Multimodal System Evaluation. (LREC 2002), Las Palmas, Spain*. Citeseer.

Kong, Y. and Y. Fu (2018). "Human action recognition and prediction: A survey". *arXiv preprint arXiv:1806.11230*.

Lee, N., W. Choi, P. Vernaza, C. B. Choy, P. H. Torr, and M. Chandraker (2017). "Desire: Distant future prediction in dynamic scenes with interacting agents". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 336–345.

Liu, H., T. Fang, T. Zhou, and L. Wang (2018). "Towards robust human-robot collaborative manufacturing: Multimodal fusion". *IEEE Access*. 6: 74762–74771.

Liu, Z., Q. Liu, W. Xu, Z. Liu, Z. Zhou, and J. Chen (2019). "Deep learning-based human motion prediction considering context awareness for human-robot collaboration in manufacturing". *Procedia CIRP*. 83: 272–278.

Monfort, M., A. Andonian, B. Zhou, K. Ramakrishnan, S. A. Bargal, Y. Yan, L. Brown, Q. Fan, D. Gutfreund, C. Vondrick, *et al.* (2019). "Moments in time dataset: one million videos for event understanding". *IEEE transactions on pattern analysis and machine intelligence*.

Oviatt, S. *et al.* (2003). "Multimodal interfaces". *The human-computer interaction handbook: Fundamentals, evolving technologies and emerging applications*. 14: 286–304.

Özyer, T., D. S. Ak, and R. Alhajj (2021). "Human action recognition approaches with video datasets—A survey". *Knowledge-Based Systems*: 106995.

Rao, R. P. and D. H. Ballard (1999). "Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects". *Nature neuroscience*. 2(1): 79.

Recupero, D. R., M. Dragoni, and V. Presutti (2015). "ESWC 15 Challenge on Concept-Level Sentiment Analysis". In: *Semantic Web Evaluation Challenges*. Ed. by F. Gandon, E. Cabrio, M. Stankovic, and A. Zimmermann. Cham: Springer International Publishing. 211–222. ISBN: 978-3-319-25518-7.

Roitberg, A., N. Somani, A. Perzylo, M. Rickert, and A. Knoll (2015). "Multimodal human activity recognition for industrial manufacturing processes in robotic workcells". In: *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*. 259–266.

Simões, B., R. De Amicis, I. Barandiaran, and J. Posada (2018). "X-reality system architecture for industry 4.0 processes". *Multimodal Technologies and Interaction*. 2(4): 72.

Simonyan, K. and A. Zisserman (2014). "Two-stream convolutional networks for action recognition in videos". In: *Advances in neural information processing systems*. 568–576.

Soomro, K., A. R. Zamir, and M. Shah (2012). "UCF101: A dataset of 101 human actions classes from videos in the wild". *arXiv preprint arXiv:1212.0402*.

Tani, J. (2016), *Exploring robotic minds: actions, symbols, and consciousness as self-organizing dynamic phenomena*. Oxford University Press.

Thomas, P. and W. David (1992). "Augmented reality: An application of heads-up display technology to manual manufacturing processes". In: *Hawaii international conference on system sciences*. 659–669.

Vélaz, Y., J. Rodríguez Arce, T. Gutiérrez, A. Lozano-Rodero, and A. Suescun (2014). "The influence of interaction technology on the learning of assembly tasks using virtual reality". *Journal of Computing and Information Science in Engineering*. 14(4).

Vicari, M. and M. Gaspari (2020). "Analysis of news sentiments using natural language processing and deep learning". *AI & SOCIETY*. DOI: 10.1007/s00146-020-01111-x. URL: https://doi.org/10.1007/s00146-020-01111-x.

Waibel, A., M. T. Vo, P. Duchnowski, and S. Manke (1996). "Multimodal interfaces". *Artificial Intelligence Review*. 10(3): 299–319.

Wang, L. (2019). "From intelligence science to intelligent manufacturing". *Engineering*. 5(4): 615–618.

Wang, P., H. Liu, L. Wang, and R. X. Gao (2018). "Deep learning-based human motion recognition for predictive context-aware human-robot collaboration". *CIRP annals*. 67(1): 17–20.

Weichert, D., P. Link, A. Stoll, S. Rüping, S. Ihlenfeldt, and S. Wrobel (2019). "A review of machine learning for the optimization of production processes". *The International Journal of Advanced Manufacturing Technology*. 104(5): 1889–1902. DOI: 10.1007/s00170-019-03988-5. URL: https://doi.org/10.1007/s00170-019-03988-5.

Xiong, Q., J. Zhang, P. Wang, D. Liu, and R. X. Gao (2020). "Transferable two-stream convolutional neural network for human action recognition". *Journal of Manufacturing Systems*. 56: 605–614.

Yuan, C., B. Wu, X. Li, W. Hu, S. Maybank, and F. Wang (2016). “Fusing $\mathcal{R}$ Features and Local Features with Context-Aware Kernels for Action ecognition”. *International Journal of Computer Vision*. 118(2): 151–171.

Yue-Hei Ng, J., M. Hausknecht, S. Vijayanarasimhan, O. Vinyals, R. Monga, and G. Toderici (2015). “Beyond short snippets: Deep networks for video classification”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 4694–4702.

DOI: 10.1561/9781680838770.ch5

Chapter 5

A Review of Explainable Artificial Intelligence in Manufacturing

By Georgios Sofianidis, Jože M. Rožanec, Dunja Mladenić and Dimosthenis Kyriazis

The implementation of Artificial Intelligence (AI) systems in the manufacturing domain enable higher production efficiency, outstanding performance, and safer operations, leveraging powerful tools such as deep learning and reinforcement learning techniques. Despite the high accuracy of these models, they are mostly considered black boxes: they are unintelligible to the human. Opaqueness affects trust in the system, a factor that is critical in the context of decision-making. We present an overview of Explainable Artificial Intelligence (XAI) techniques as a means of boosting the transparency of models. We analyze different metrics to evaluate these techniques and describe several application scenarios in the manufacturing domain.

5.1 Introduction

The increasing digitalization of every aspect of life provides vast amounts of data, enabling the implementation of Artificial Intelligence (AI) models. The manufacturing and process industry is not an exception to this trend. AI models play a significant role in many aspects of the manufacturing process. AI models drive better quality by enhancing quality inspection and process monitoring in production lines, ease reconfiguration and customization of automated part handling, fault diagnosis and event prediction, more agile production management, flexible production planning, and enabling safe collaboration between humans and cobots. Especially the latter is a big step towards the transition into Industry 5.0, where the focus is on the synergy between humans and robots and the actors are collaborators instead of competitors.

AI models provide the means to automate many tasks and achieve unprecedented performance levels. However, in most cases, such models are opaque to the user: they work as black-boxes. Their predictions are mostly accurate, but no intuition behind the reasoning process is available to human users. Given the impact of those predictions on the decision-making processes, it is crucial to develop mechanisms and techniques to provide insights to users on such an AI model reasoning process. The development of such techniques and mechanisms and how those insights are presented has given birth to a research field of its own, known as Explainable Artificial Intelligence (XAI). While the field of XAI can be traced back to the 1970's [44], it has experienced a new flourishment since the rise of modern deep learning [55].

Though there is no single definition of the scope of this research field, most authors agree it includes intrinsically interpretable models and post-hoc explainability models (the model's capability of being explained by another interpretable model). Authors identify two sources of model opacity (or opaqueness) [5]: (i) the complexity of the formal structure of the model is beyond human comprehension, or alien to human reasoning, or (ii) because the inner workings of the model cannot be shared (e.g., being considered a trade secret). Model opaqueness can be relative to expert knowledge: e.g., it can be opaque to an analyst but not to the machine learning engineer. [32] introduced the term *deep opacity* to describe models whose opacity cannot be removed even by human experts. When presenting insights on the reasoning process of an AI model, the explanations should resemble a logic explanation [43], and take into account relevant context. [19] considers context has three elements related to the explainee: (i) *Profile* (user profile, to whom we present the explanation), (ii) *Objective* (refer to the goals of the explanation, e.g., are the explanations meant to improve the model, enhance trust in the system, aid on decision-making or foster action based on decisions made), and (iii) *focus* (if the explanation is either global or local). In local explanations, the specific point of

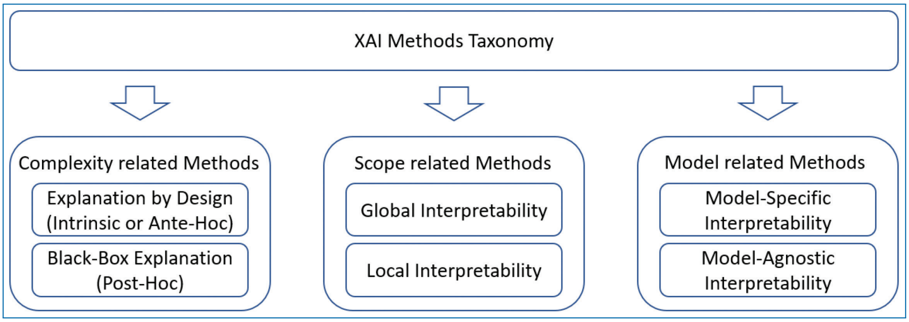

XAI Methods Taxonomy
Complexity related Methods
Explanation by Design (Intrinsic or Ante-Hoc)
Black-Box Explanation (Post-Hoc)
Scope related Methods
Global Interpretability
Local Interpretability
Model related Methods
Model-Specific Interpretability
Model-Agnostic Interpretability

Figure 5.1. XAI taxonomy.

interest must be considered part of the context. When the explanations aim to aid decision-making or take action, they should provide information regarding actionable features.

XAI techniques and methods can be classified into three categories, considering the explainability source, the scope of the explanation, and the level of dependency on the forecasting model used (see Fig. 5.1). We distinguish intrinsically explainable models and forecasting models that require post-hoc models to get insights into the forecast's reasoning process regarding the explainability source. Concerning the explanation's scope, explanations can be global (describe the behavior of the whole model for the average of forecasts provided) or local (describe the model's behavior for a particular forecast). Finally, regarding the dependency on the forecasting model's explanation, we distinguish model-agnostic (can be applied to any AI model) or model-specific techniques (can be applied only to AI models built with a particular algorithm or type of algorithms).

In this chapter, we introduce the field of Explainable Artificial Intelligence, describing methods and techniques used to identify meaningful features driving forecasts, current approaches used to evaluate such models, applications and use cases in the industrial domain, and open challenges. When doing so, we do not consider intrinsically explainable models.

5.2 Methods and Techniques

Different methods and techniques have been introduced to boost the transparency and acceptance of AI models and different taxonomies have been proposed in literature based on the explanation generating mechanism, the type of explanation, the scope of explanation, the type of model it can explain, or a combination of these features. [1] classified those methods into intrinsic interpretable models and post-hoc explanations and divided the latter to text explanations, visual explanations, local explanations, explanations by example, explanations by simplification, and feature relevance explanations techniques. [4] introduced a categorization of explanation methods based on the type of explanation returned and divided them based on the most common data types such as tabular, image, and text. For tabular data, feature importance is one of the most popular types of explanation returned by local explanation methods. The explainer assigns to each feature an importance value which represents how much that particular feature was important for the prediction under analysis. The sign and magnitude of each importance value are also considered to understand the contribution of each feature. Similar to the above but in the field of image classification, saliency maps can be used as explanations. Those are modeled as matrices with the same dimensions as that of the image we want to explain, and each element of the matrix represents the saliency of each pixel to the forecast. Another type of explanation that can be implemented on tabular data is the rule-based explanation. Human readable decision rules can give the end-user an explanation about the reasons that lead to the final prediction. A decision or factual or logic rule is a set of premises that lead to a specific forecast. Counterfactual rules are a set of rules that lead to the opposite of a specific forecast. [30] classified XAI techniques according to the type of explanation and the scope of explanation. The three types he distinguished are model-based, attribution-based, and example-based explanations. In this chapter, we present some of the well-known explainability methods based on the taxonomy introduced by [30].

The class of *model-based explanations* include methods that are either explainable by nature (intrinsic explainability) or methods that use a different interpretable model to explain the task model (post-hoc explainability). The first subclass can be divided into sparse linear classifiers (e.g., linear or logistic regression, generalized additive models (GAMs)), discretization methods (e.g., rule-based learners, decision trees), and example-based models (e.g., K-nearest neighbors). The second subclass includes interpretable surrogate models that can approximate the task model and can be used as post-hoc explanations.

The class of *attribution-based explanations* use the explanatory power of input features to explain the task model. These approaches are also known as feature (a.k.a variable) importance, relevance, or influence methods. Most post-hoc explanations fall under this category which can further be divided into perturbation-based and backpropagation-based methods.

Among the perturbation-based methods, we can find the *Prediction Difference Analysis (PDA)* [40], which is based on the idea that the relevance of an input feature concerning the class can be estimated by measuring how the predictions change if this particular feature is removed. This method cannot deal with saturated classifiers (models whose output does not change after removing part of the features). A similar approach for images was developed by [60] with the *Deconvolutional Networks*, which attempts to reconstruct the feature map into the layer input or the original image. The proposed networks used convolution, max-pooling layers, and the ReLU activation function. Sliding a gray-color square over the image, they measure changes in feature activations and the classification scores. A variation of this method was developed by [11], who, instead of using a gray-square, replaces regions of an image with constant values, noise, or performs some blurring on the image. This method was evolved by [35], who chose upsampled, random binary masks to perform the occlusions and analyzed their impact on the target class classification score. Another variation of [60] was introduced by [63], who removed several features at once by using prior knowledge about images and choosing patches of connected pixels as feature sets to analyze the effects of different window sizes on top scoring classes. The huge computational cost of this method was later minimized by [13] through the *Contextual Prediction Difference Analysis*, which also solved the problem of saturated classifiers by producing a model-aware saliency map.

Another family of explainability methods computes feature attributions from a forward or backward pass through the network. They require architectural or back-propagation rule modifications or access to intermediate layers. However, most of these methods have lower computational costs than the ones mentioned above, leading to faster results. One of the first approaches of this kind was introduced by [47], who computed feature attributions by taking the partial derivative of the output class with respect to the input. The resulting absolute values allow identifying which input features can be perturbed the least for the output to change the most. A drawback of this method is that it is noisy, and the absolute value of the gradients prevents the detection of positive and negative evidence in the input. This approach was improved by the *Gradient * Input* method [46], which increases the sharpness of attribution maps by taking the signed partial derivatives of the output with respect to the input and multiplying feature-wise by the input itself. The multiplication with the input indicates the interest in the salience rather than sensitivity. [46] introduced the *Deep Learning Important FeaTures (DeepLIFT)* method, which uses a derivative-based method to propagate activation differences instead of gradients through the network. The intuition behind the method is that though the partial derivatives do not explain a single decision, they indicate what change in the image could make a change in the prediction. In the same line, [53] developed the *Integrated Gradients* approach, which relies on the idea of computing attributions by multiplying the input variable element-wise with the average partial derivative, as the input varies from a baseline to its final value. *Smooth-Grad* [49]takes a different approach, and focuses on local sensitivity, and calculates averaging maps with a smoothing effect made from several small perturbations of an input image. The effect is enhanced by further training with these noisy images. Finally, it sharpens the sensitivity maps, to increase their quality. [60] was evolved by [52], who proposed the *All Convolutional Net*, as an alternative that replaces the max-pooling layer for convolutional layers with an increased stride. A slightly different approach was proposed by [61], who introduced the *Class Activation Mapping (CAM)*. This method relies on the observation that some convolutional layers behave as unsupervised object detectors, and it uses global average pooling to create heat maps of a pre-softmax layer. The heat maps point out the regions of an image that are responsible for a prediction. *Gradient-weighted Class Activation Mapping (GradCAM)* [45] uses the gradient information to understand how strongly does each neuron activate in the last convolutional layer of the neural network. The localizations are combined with existing high-resolution visualizations to obtain high-resolution class-discriminative guided visualizations as saliency masks. The CAM and GradCAM approaches inspired the *GradCAM++* method [6], which combines the positive partial derivatives of feature maps of a rear convolutional layer with a weighted special class score to explain the occurrence of multiple object instances in an image. *Layer Wise Relevance Propagation (LRP)* [3] is a gradient method suffering from vanishing gradient problems. The main idea behind this is the decomposition of the prediction function as a sum of layer-wise relevance values. The prediction is redistributed backward using local redistribution rules until assigning a relevance score to each input feature. There are different variations of the LRP algorithm based on the backward redistribution rule.

Many explainability methods were built, relying on surrogate models to provide explanations regarding the reference model. One of such methods is *TREPAN* [7] which provides heuristics to issue queries against neural networks and create a decision tree that approximates forecasts from the given network, while providing an interpretable set of rules that explain the forecast. A more general approach was presented in the *Local Interpretable Model-agnostic Explanations (LIME)* [38], which can explain the predictions of any AI model through a post-hoc, local, linear, and interpretable model. The model attempts to learn a particular forecast, by matching the given feature vector and perturbed inputs, to the results obtained from the reference model. Since the creation of LIME, multiple variants were developed. *k-LIME* ([16]) uses local generalized linear model surrogates to explain the predictions, while local regions are defined by k clusters instead of perturbed samples. The criteria to define the value of k is to K is that predictions from the local generalized linear models maximize R^2. In addition to this, a global surrogate linear generalized model is trained to provide information about overall feature average trends. *DLIME* ([58]) proposes a deterministic version of LIME, where instead of random perturbations, they apply agglomerative hierarchical clustering to group the training data. The hierarchical clustering does not require prior knowledge regarding clusters. A dendrogram is cut where the gap is the largest between two successive groups to determine the number of clusters. A k-Nearest Neighbour classifier is trained to classify new instances into those clusters based on the clusters obtained. All data points belonging to a given cluster are used to train a linear model, which provides deterministic and consistent local explanations. *LIMEtree* ([50]) follows a similar approach to LIME, building a regression tree as surrogate model. The regression tree enables capturing non-linear relationships between the interpretable features and the target variable. At the same time, it does not require independence between interpretable features. The authors consider the model's biggest advantage is providing personalized counterfactual explanations through an interactive interface that enables imposing certain conditions on the sample of interest. Inspired in LIME, [9] developed STREAK, an interpretability method for neural networks conceived as a set function maximization, achieving similar accuracy than LIME, while having a faster runtime execution. A slightly different approach is presented in Anchors [39], where a set of rules replaces the surrogate model. Since the local behavior of a model can be highly non-linear, the authors propose using a set of if-then rules, which are intuitive and easy to understand. To explore the model's behavior in the perturbation space, the authors apply multi-armed bandits to incrementally construct the rules, generate candidate predicates, and choose the one with the highest precision until a given precision threshold is reached with a high probability. *LoRE - Local Rule-Based Explanations* [14] proposes a parameter-free, two step method that also provides rule-based explanations. First, it creates a balanced set of neighbor instances using a genetic algorithm to explore the decision boundary of the data point of interest. Then it builds a decision tree classifier, which enables to derive decision rules and counterfactuals. *Local Foil Trees* [54] specifically deal with generating counterfactual explanations. To that end, they consider two possible outputs: the model forecast (fact), and the desired label (foil). A decision tree is then built based on the local dataset. The rules are computed from the difference between paths regarding the *"fact leaf"*, and *"foil leaf"*.

While most explainability methods based on surrogate models provide specific techniques, [17] developed a framework that enabled comparing surrogate models on three dimensions: data sampling, explanation generation, and interaction. [51] considered a slightly different approach and developed an algorithmic framework (*bLIMEy – build LIME yourself*) that enables building custom local surrogate explainers for model predictions, considering three dimensions: data sampling, explanation generation, and interpretable representation.

Another local-agnostic explanation method is *SHAP* [28] which stands for SHapley Additive exPlanations and can be used to produce several explanation models. These models compute SHAP values: a unified measure of feature importance based on the Shapley values, a concept from cooperative game theory. The different explanation models proposed by SHAP differ on how they approximate the computation of the SHAP values. The explanation models provided by SHAP are called *additive feature attribution methods*. The construction of the SHAP values allows to employ them both locally, in which each observation gets its own set of SHAP values, and globally, by exploiting collective SHAP values.

In the image classification field, two explanators can be implemented for deep networks: DEEP-SHAP and GRAD-SHAP. DEEP-SHAP is a high-speed approximation algorithm for shap values in deep learning models that connect with the DeepLift algorithm. The implementation is different from the original DeepLift by using a baseline distribution of background samples instead of a single value and using Shapley equations to linearise non-linear components of the black-box such as max, softmax, products, divisions. GRAD-SHAP, instead, is based on IntGrad and SmoothGrad algorithms. IntGrad values are a bit different from SHAP values, and require a single reference value to integrate from. As an adaptation to approximate SHAP values, GRAD-SHAP reformulates the integral as an expectation and combines that expectation with sampling reference values from the background dataset as done in SmoothGrad.

Another family of explainability techniques is that of *example-based explanations*. Methods in this class explain the task model by selecting particular instances from the dataset that describe the model or by creating new instances. Instances that are well predicted by the forecasting model (prototypes) and instances that are not well predicted by the model (criticism) are the influential instances for the model parameters or output, while counterfactual explanations indicate the required changes in the input side that will have significant changes (e.g., reverse the prediction) in the prediction/output. [21] proposed a methodology named *MMD-CRITIC* to learn prototypes and criticisms for a given dataset using the maximum mean discrepancy (MMD) as a measure of similarity. [36] introduced *MAPLE*. This post-hoc local agnostic explanation method can also be used as a transparent model due to its internal structure. It combines random forests with feature selection methods to return feature importance-based explanations. *DICE* which stands for Diverse Counterfactual Explanations [31] is a local, post-hoc and agnostic method that solves an optimization problem with several constraints to ensure feasibility and diversity when returning counterfactuals. Feasibility is critical in the context of counterfactuals since it allows avoiding examples that are unfeasible.

We classify the aforementioned methods according to multiple criteria in Table 5.1.

5.3 Evaluation Measures

Explainability is considered a subjective concept. [30] considers that an AI system is explainable if either the model is intrinsically interpretable or if the

Table 5.1. Classification of XAI techniques.

Explanation Technique	Reference	Model Based	Attribution Based	Example Based	Local (L)/ Global (G)	Agnostic (A)/ Specific (S)	Data Type
All Convolutional Net	[52]	X	X		L	S	IMAGE
Anchors	[39]		X		L/G	A	TABULAR/ TEXT
Class Activation Mapping (CAM)	[61]		X		L	S	IMAGE
Contextual Prediction Difference Analysis	[11]		X		L	S	IMAGE
Deconvolutional Networks	[60]	X	X		L	S	IMAGE
Deep Learning Important FeaTures (DeepLIFT)	[46]		X		L	S	ANY
DICE	[31]			X	L	A	ANY
DLIME	[58]	X	X		L	A	ANY
GradCAM++	[6]		X		L	S	IMAGE
Gradient	[47]		X		L	S	ANY
Gradient * Input	[46]		X		L	S	ANY
Gradient Weighted Class Activation Mapping (GradCAM)	[45]		X		L	S	IMAGE
Integrated Gradients	[53]		X		L	S	ANY
k-LIME	[16]	X	X		L	A	ANY
Layer Wise Relevance Propagation (LRP)	[3]		X		L	A	ANY
LIME	[38]	X	X		L	A	ANY
LIMETree	[50]	X	X		L	A	TAB
Local Foil Trees	[54]	X		X	L	A	TABULAR
LoRE	[14]		X		L	A	TABULAR
MAPLE	[36]	X	X		L	A	TABULAR
Meaningfull Perturbation	[11]		X		L	S	IMAGE
MMD-CRITIC	[21]			X	G	A	ANY
Prediction Difference Analysis (PDA)	[40, 63]		X		L	S	IMAGE
RISE	[35]		X		L	S	IMAGE
SHAP	[28]		X		L/G	A	ANY
Smooth Grad	[49]		X		L	S	IMAGE
STREAK	[9]				L	A	IMAGE
TREPAN	[7]		X		G	S	TABULAR

non-interpretable model can be complemented with an interpretable and faithful explanation. While the XAI techniques provide different kinds of information, the perceived quality of the explanations depends on the users, the domain, the information of interest, and the explanation itself. To evaluate the explanations, it is necessary to define different criteria of goodness for an explanation. Given an interpretable approximation for a reference, model [25] lists four aspects to be considered on evaluation: fidelity (ability to capture the reference model behavior correctly), unambiguity (ability to provide a single and deterministic rationale to explain each data instance), interpretability (the approximation should be human-understandable), and interactivity. The aspect of fidelity is further elaborated by [22], who considers two properties: soundness (the extent to which each explanation component is truthful to the reference model) and completeness (the extent to which the explanation describes the reference model). [56] enumerate another three criteria: sensitivity, the degree of integration, and cognitive salience. Sensitivity is defined as the strength of the relationship of explanatory variables with background conditions: the weaker the relationship, the more convincing the explanation. The degree of integration refers to the connectedness of the explanation to a larger theoretical framework. Finally, cognitive salience is defined as the ease with which the rationale behind the explanation can be followed.

The aforementioned criteria require different evaluation approaches. [8] identified three categories of them:

- **Application-grounded evaluation:** grounded in a real-world application, collects domain expert's feedback regarding the explanations provided to them.
- **Human-grounded evaluation:** refers to feedback obtained from experiments performed with lay users, when no real-world application exists in place.
- **Functionality-grounded evaluation:** the evaluation is performed considering some formal definition or criteria, that measures the explanation quality.

To assess the explainability methods, [15] propose three tests for functionality-grounded evaluations: **Feature Augmentation Test**, **Synthetic Test**, and **Feature Deduction Test**. The **Feature Augmentation Test** considers that if the values of the explainable features from a specific instance are replaced by the values of those features from an instance with a different label (e.g., "new-label"), the classification outcome should be "new-label". The **Synthetic Test** is based on the assumption that if the explainability features are accurately selected, new synthetic instances can be created by preserving the explainability feature values and assigning random values to the rest of the features without affecting the forecast outcome. Finally, the **Feature Deduction Test** considers that if the selected explainability features are correctly selected, removing one of them from the input should lead to a different forecast. Even though this approach is frequently adopted in the literature [11, 20, 35, 60, 63] pointed out that samples, where a subset of features are removed have a different data distribution than the samples the model was trained on, violating a key machine learning assumption. They instead propose the RemOve And Retrain (ROAR) approach, which for each feature deemed important, they replace it by a non-informative value in the train and test sets, retrain the model and measure the performance change. In addition to this technique, they propose using a random assignment of feature importance as a benchmark to measure the quality of explainability feature extraction techniques.

There is currently little research regarding application and human-grounded evaluations [8, 62]. A popular and domain-specific method is to evaluate to create a heatmap regarding model sensitivity to region-based perturbations. According to the heatmap, the main idea behind this is that the perturbation of relevant input variables would lead to a decline in prediction score than the perturbation of input features with less importance. [22] used questionnaires with short responses and Likert scales. In contrast, [23] used three quantitative metrics: accuracy, response time, and subjective satisfaction. The authors measured accuracy and response time regarding the subject response to different tasks proposed in their research. Subjective satisfaction was measured on a Likert scale for each explanation. [24] proposed the Human Interpretability Score (HIS – see Eq. 5.1), which constitutes an alternative metric regarding the user's response time. On the other side, there is a wider set of metrics reported for functionality-grounded evaluations.

$$HIS(x, R) = \begin{cases} 0, & \text{if } RT_{mean}(x, R) > RT_{max} \\ RT_{max} - RT_{mean}(x, M), & RT_{mean}(x, R) \leq RT_{max} \end{cases} \tag{5.1}$$

Equation 5.1: Human Interpretability Score. Measures how long it takes the user to predict the label assigned to certain data point, assigning a cap to the response time. x and R correspond to the instance and model considered.

Among the metrics proposed by [33] we find *Mutual Information*, *Diversity*, *Monotonicity*, *Non-sensitivity*, and *Effective complexity*. *Mutual Information* is considered when creating an interpretable data representation. [33] proposes measuring Mutual Information on two cases: (i) between the features of the original model and the subset of explainable features, and (ii) against the target values. Ideally, the number of explainable features should be reduced to maximize simplicity and broadness, while aiming towards keeping a high fidelity regarding the target label (see Eq. 5.2).

$$I(x, y) = D_{KL}(P_{(x,y)} \parallel P_x \otimes P_y) \tag{5.2}$$

Equation 5.2: Mutual Information. Measures the mutual dependence between two random variables x and y.

Diversity attempts to measure the degree to which a set of rules integrates to the explanation (see Eq. 5.3). **Monotonicity** considers that feature attributions should be monotonic. [33] proposes measuring it as the Spearman's correlation between two vectors: (i) the absolute values of attributions, and (ii) the corresponding expectations. The intuition behind the **Non-sensitivity** metric (see Eq. 5.4) is to assess that the explainability method does not assign any relevance score to the features the model is not functionally dependent on. The authors compute it as the cardinality of the symmetric difference between features assigned zero attribution and the features the model does not functionally depend on. **Effective complexity** measures if some explanation features can be ignored without significantly affecting the prediction (see Eq. 5.5).

$$Diversity = \sum_{x_i, x_j \in E; x_i \neq x_j} \frac{d(x_i, x_j)}{2N_{\mathrm{E}}} \tag{5.3}$$

Equation 5.3: Diversity metric. E is the set of examples considered, d is a distance metric for the space X, while N_E corresponds to the number of examples.

$$|A_0 \triangle X_0| \tag{5.4}$$

Equation 5.4: Non-sensitivity. A_0 represents featues with zero attribution, X_0 refers to features on which the model is not functionally dependent on. $|\cdot|$ denotes the set cardinality, and $\triangle$ the symmetric set difference.

$$k* = argmin_{k \in 1,\ldots,N} |M_k| \ where \ E(l(y*, f - M_k)|x*_{M_k}) < \varepsilon \tag{5.5}$$

Equation 5.5: Effective Complexity. M_k denotes the set of top k features, x denotes features, $\varepsilon > 0$ corresponds to some arbitrary tolerance, $f - M_k$ is the restriction of the model R to non-important features, given M_k.

The **Local Approximation Accuracy** was proposed by [15] to compare the decision boundary of the surrogate model against the original one. The authors do so by computing the Root Mean Squared Error between the original and surrogate model predictions on the test samples. A similar intuition is present in the **Disagreement** metric proposed by [25]. For a classification setting, they attempt to measure the surrogate model fidelity by computing the disagreement between labels of the surrogate model and the original one (see Eq. 5.6).

$$Disagreement(R) = \sum_{i=1}^{N} \Big| x \Big| x \in D, x \ satisfies \ q_i \wedge s_i, B(x) \neq c_i \Big| \tag{5.6}$$

Equation 5.6: Disagreement metric. Quantifies the disagreement between a surrogate model R and the reference forecasting model *B*, given a dataset *D*. The triplet *(q, s, c)* stands for (feature, operator, class).

[25] propose another six metrics to evaluate forecast explanations: rule overlap, cover, the rule set size (see Eq. 5.7), the rule set maximum width, the number of descriptor sets, and feature overlap. The **Rule overlap** computes the overlap between pairs of rules defined in the surrogate model. It is expected that the lower the overlap, the lower the surrogate model ambiguity (see Eq. 5.8). **Cover** is defined as the number of instances that match a given rule from the surrogate model (see Eq. 5.9). The **Maximum Width** refers to the maximum width obtained from computing the width over all the elements from the surrogate model. The authors define an element as either rule conditions or neighborhood descriptors (see Eq. 5.10). The authors define the **Number of Unique Descriptor Sets** as the number of unique neighborhood descriptors provided in the surrogate model (see Eq. 5.11). Finally, the **Feature overlap** measures the features overlap between every pair of unique neighborhood descriptor and rule (see Eq. 5.12).

$$RuleSetSize(R) = NumberOfRules(q, s, c) \tag{5.7}$$

Equation 5.7: Rule set size. *R* denotes the decision set. The triplet *(q, s, c)* stands for (feature, operator, class). The triplets are contained in the decision set.

$$RuleOverlap(R) = \sum_{i=1}^{N} \sum_{j=1, j \neq i}^{N} overlap(q_i \wedge s_i, q_j \wedge s_j) \tag{5.8}$$

Equation 5.8: Rule overlap. *R* denotes the decision set. The triplet *(q, s, c)* stands for (feature, operator, value).

$$cover(R) = \Big| x \Big| x \in D, x \; satisfies \; q_i \wedge s_i, \;\; where \; i \in 1 \ldots N \Big| \tag{5.9}$$

Equation 5.9: Cover. *R* denotes the decision set. The triplet *(q, s, c)* stands for (feature, operator, value). *D* represents a dataset, and *x* and instance in such dataset.

$$MaximumWidth(R) = max(width(e)), e \in \bigcup_{i=1}^{N} (q_i \cup s_i) \tag{5.10}$$

Equation 5.10: Maximum Width. *R* denotes the decision set. *e* represents elements, which can be ether rule conditions or neighborhood descriptors.

$$NumberOfUniqueDescriptorSets(R) = |dset(R)|, \;\; where \; dset(R) = \bigcup_{i=1}^{N} (q_i) \tag{5.11}$$

Equation 5.11: Number of Unique Descriptor Sets. R denotes the decision set, and q denotes features.

$$FeatureOverlap(R) = \sum_{i=1}^{N} FeatureOverlap(q, s_i) \quad (5.12)$$

Equation 5.12: Feature Overlap. R denotes the decision set, q denotes features in descriptor sets, and s denotes operators.

A different set of metrics is considered by [37], who for tree-based models measured the mean path length, the mean number of distinct features in a path, the number of nodes, and the number of nonzero features. Finally, [48] reported assessing explainability methods based on the total number of runtime operation counts performed by the model when computing the forecast for a given input.

5.4 Applications, use Cases and Open Issues

Though multiple XAI methods exist, they do not suffice by themselves to provide human-understandable explanations. They are built into frameworks and applications that provide a convenient interface and additional context to achieve that goal. One such framework is bLIMEy [51], which decomposes surrogate models into three steps: interpretable data representation (transform data from the original to the interpretable domain), data sampling, and explanation generation. [18] follows a similar approach and describes the IBEX (Interactive Black-box EXplanation system) framework with two components: an explainer that produces explanations based on user's needs, and a sampling component, that selects appropriate inputs to create the explanation. [2] describes *AI Explainability 360*, an extensible toolkit developed that provides contextual explainers based on the stage of the AI model development pipeline, kind of model, and explanation requirements. [34] explores the usage of domain knowledge encoded in an ontology improves the quality of the explanations. [42] explores the usage of semantic technologies to abstract relevant concepts encoded in the features, avoid exposing sensitive details regarding the forecasting model, and provide higher-level information to the users. The authors complement model explanations with information regarding real-world events reported in the media that likely influenced the variables of interest. [41] developed an ontology to model user's feedback based on a given forecast and provided explanations. [59] developed an intelligent assistant for manufacturing, which creates directive explanations for the users using heuristics and domain knowledge. The application tracks user's implicit and explicit feedback regarding local forecast explanations, enabling application-grounded evaluations.

The integration of explainability methods into applications enables providing relevant information regarding model forecasts to different stakeholders. For instance, data scientists and machine learning engineers require low-level data to monitor the AI model behavior, identify corner cases, and work towards a more accurate and robust model. On the other side, employees and supervisors require high-level insights that convey reasons behind the model forecasts, can interactively explore different *"what-if"* scenarios, and provide feedback regarding the explanations provided. We envision explainability methods can be useful in a wide range of manufacturing use cases, such as automatic defect detection (inform the user on the image regions influencing the decision), production planning (provide an insight on the cost of the opportunity given different scheduling decisions), or demand forecasting (provide insights why we expect demand will take place and which factors affect the quantity estimates).

Several explainability techniques have been implemented in the manufacturing domain and specifically the predictive quality management domain (Quality 4.0) to boost the transparency of AI deployed models. [12] used XAI techniques such as CAM and Contrastive gradient-based saliency maps to explain black-box classifiers in the area of quality welds in ultrasonically welded battery tabs. They produced heatmaps where they visualized several color maps to gain insights into true positive versus false-positive predictions. [27] implemented several XAI methods to provide explanations for domain experts in the area of defect classification of thin-film-transistor liquid-crystal display panels. Techniques such as CAM, LRP, integrated gradients, guided backpropagation, and SmoothGrad were implemented and visualized on a VGG-16 classification model. Based on the visualized results, LRP and guided backpropagation were selected as they produced well-distributed heatmaps. Moreover, by fitting the model into a decision tree and converting the prediction results into human interpretable text, the authors achieved the maximum level of explainability when they presented the results to domain experts for evaluation purposes. In the area of manufacturing cost estimation, [57] described a method based on visualization of the machining features of a 3D computer aided design model that are influencing the increase in manufacturing costs. For the proposed purpose, a 3D gradient-weighted class activation mapping as XAI method was applied.

Cybersecurity in a transversal concern related to all smart manufacturing cases. XAI techniques were successfully applied in the cybersecurity domain, to support the exploration of model vulnerabilities [26, 29], and identify perturbed data samples [10].

In the European Horizon 2020 project STAR (Safe and Trusted Human Centric Artificial Intelligence in Future Manufacturing Lines), XAI is used to provide insights on most relevant features to each forecast, explore model vulnerabilities and help identify potential data poisoning. While providing accurate explanations to forecasts provides the users additional elements for decision-making, the vulnerabilities assessment and early data poisoning identification ensures the system is secure, enhancing users trust in the system.

5.5 Conclusion

The new industrial revolution relies on AI to enable higher production efficiency, and safer operations. XAI techniques provide means to reduce black-box models opaqueness, and increase trust in the system. In this contribution, we introduce the field of XAI. We list several taxonomies found in the literature alongside state-of-the-art methods and techniques to interpret AI models. We also include metrics with different qualitative and quantitative characteristics as a means of evaluating the above methods. Finally, we list applications of XAI, describe several use cases in the manufacturing domain, and open opportunities.

XAI requires a multi-disciplinary approach. Special consideration needs to be given to understand how domain experts and end-users operate. Users must be involved in the XAI outcomes validation. The integration of XAI into manufacturing processes will be paramount for the transition into the fifth industrial revolution.

Acknowledgements

This work has been carried out in the H2020 STAR project, which has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 956573.

References

[1] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-López, Daniel Molina, Richard Benjamins, *et al.* Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information Fusion*, 58:82–115, 2020.

[2] Vijay Arya, Rachel KE Bellamy, Pin-Yu Chen, Amit Dhurandhar, Michael Hind, Samuel C Hoffman, Stephanie Houde, Q Vera Liao, Ronny Luss, Aleksandra Mojsilovic, *et al.* Ai explainability 360: An extensible toolkit for understanding data and machine learning models. *Journal of Machine Learning Research*, 21(130):1–6, 2020.

[3] Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one*, 10(7):e0130140, 2015.

[4] Francesco Bodria, Fosca Giannotti, Riccardo Guidotti, Francesca Naretto, Dino Pedreschi, and Salvatore Rinzivillo. Benchmarking and survey of explanation methods for black box models. *arXiv preprint arXiv:2102.13076*, 2021.

[5] Lok Chan. Explainable ai as epistemic representation. *Overcoming Opacity in Machine Learning*, page 7, 2021.

[6] Aditya Chattopadhay, Anirban Sarkar, Prantik Howlader, and Vineeth N. Balasubramanian. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 839–847. IEEE, 2018.

[7] Mark Craven and Jude Shavlik. Extracting tree-structured representations of trained networks. *Advances in neural information processing systems*, 8:24–30, 1995.

[8] Finale Doshi-Velez and Been Kim. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*, 2017.

[9] Ethan R. Elenberg, Alexandros G. Dimakis, Moran Feldman, and Amin Karbasi. Streaming weak submodularity: Interpreting neural networks on the fly. *arXiv preprint arXiv:1703.02647*, 2017.

[10] Gil Fidel, Ron Bitton, and Asaf Shabtai. When explainability meets adversarial learning: Detecting adversarial examples using shap signatures. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2020.

[11] Ruth C. Fong and Andrea Vedaldi. Interpretable explanations of black boxes by meaningful perturbation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3429–3437, 2017.

[12] Claudia V. Goldman, Michael Baltaxe, Debejyo Chakraborty, and Jorge Arinez. Explaining learning models in manufacturing processes. *Procedia Computer Science*, 180:259–268, 2021.

[13] Jindong Gu and Volker Tresp. Contextual prediction difference analysis for explaining individual image classifications. *arXiv preprint arXiv:1910.09086*, 2019.

[14] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Dino Pedreschi, Franco Turini, and Fosca Giannotti. Local rule-based explanations of black box decision systems. *arXiv preprint arXiv:1805.10820*, 2018.

[15] Wenbo Guo, Dongliang Mu, Jun Xu, Purui Su, Gang Wang, and Xinyu Xing. Lemna: Explaining deep learning based security applications. In *proceedings of*

the 2018 ACM SIGSAC conference on computer and communications security, pages 364–379, 2018.

[16] Patrick Hall, Navdeep Gill, Megan Kurka, and Wen Phan. Machine learning interpretability with h2o driverless ai. *H2O. ai*. URL: http://docs.h2o.ai/driverless-ai/latest-stable/docs/booklets/MLIBooklet.pdf, 2017.

[17] Clement Henin and Daniel Le Métayer. Towards a generic framework for black-box explanations of algorithmic decision systems. In *IJCAI 2019 Workshop on Explainable Artificial Intelligence (XAI)*, 2019.

[18] Clément Henin and Daniel Le Métayer. A generic framework for black-box explanations. In *2020 IEEE International Conference on Big Data (Big Data)*, pages 3667–3676. IEEE, 2020.

[19] Clément Henin and Daniel Le Métayer. A multi-layered approach for tailored black-box explanations. 2021.

[20] Sara Hooker, Dumitru Erhan, Pieter-Jan Kindermans, and Been Kim. A benchmark for interpretability methods in deep neural networks. *arXiv preprint arXiv:1806.10758*, 2018.

[21] Been Kim, Oluwasanmi Koyejo, Rajiv Khanna, *et al.* Examples are not enough, learn to criticize! criticism for interpretability. In *NIPS*, pages 2280–2288, 2016.

[22] Todd Kulesza, Simone Stumpf, Margaret Burnett, Sherry Yang, Irwin Kwan, and Weng-Keen Wong. Too much, too little, or just right? ways explanations impact end users' mental models. In *2013 IEEE Symposium on visual languages and human centric computing*, pages 3–10. IEEE, 2013.

[23] Isaac Lage, Emily Chen, Jeffrey He, Menaka Narayanan, Been Kim, Sam Gershman, and Finale Doshi-Velez. An evaluation of the human-interpretability of explanation. *arXiv preprint arXiv:1902.00006*, 2019.

[24] Isaac Lage, Andrew Slavin Ross, Been Kim, Samuel J. Gershman, and Finale Doshi-Velez. Human-in-the-loop interpretability prior. *arXiv preprint arXiv:1805.11571*, 2018.

[25] Himabindu Lakkaraju, Ece Kamar, Rich Caruana, and Jure Leskovec. Interpretable & explorable approximations of black box models. *arXiv preprint arXiv:1707.01154*, 2017.

[26] Sebastian Lapuschkin, Stephan Wäldchen, Alexander Binder, Grégoire Montavon, Wojciech Samek, and Klaus-Robert Müller. Unmasking clever hans predictors and assessing what machines really learn. *Nature communications*, 10(1):1–8, 2019.

[27] Minyoung Lee, Joohyoung Jeon, and Hongchul Lee. Explainable ai for domain experts: a post hoc analysis of deep learning for defect classification of tft–lcd panels. *Journal of Intelligent Manufacturing*, pages 1–13, 2021.

[28] Scott Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *arXiv preprint arXiv:1705.07874*, 2017.

[29] Yuxin Ma, Tiankai Xie, Jundong Li, and Ross Maciejewski. Explaining vulnerabilities to adversarial machine learning through visual analytics. *IEEE transactions on visualization and computer graphics*, 26(1):1075–1085, 2019.

[30] Aniek F. Markus, Jan A. Kors, and Peter R. Rijnbeek. The role of explainability in creating trustworthy artificial intelligence for health care: a comprehensive survey of the terminology, design choices, and evaluation strategies. *Journal of Biomedical Informatics*, page 103655, 2020.

[31] Ramaravind K. Mothilal, Amit Sharma, and Chenhao Tan. Explaining machine learning classifiers through diverse counterfactual explanations. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 607–617, 2020.

[32] Vincent C. Müller. Deep opacity undermines data protection and explainable artificial intelligence. *Overcoming Opacity in Machine Learning*, page 18, 2021.

[33] An-phi Nguyen and María Rodríguez Martínez. On quantitative aspects of model interpretability. *arXiv preprint arXiv:2007.07584*, 2020.

[34] Cecilia Panigutti, Alan Perotti, and Dino Pedreschi. Doctor xai: an ontology-based approach to black-box sequential data classification explanations. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, pages 629–639, 2020.

[35] Vitali Petsiuk, Abir Das, and Kate Saenko. Rise: Randomized input sampling for explanation of black-box models. *arXiv preprint arXiv:1806.07421*, 2018.

[36] Gregory Plumb, Denali Molitor, and Ameet Talwalkar. Model agnostic supervised local explanations. *arXiv preprint arXiv:1807.02910*, 2018.

[37] Forough Poursabzi-Sangdeh, Daniel G. Goldstein, Jake M. Hofman, Jennifer Wortman Vaughan, and Hanna Wallach. Manipulating and measuring model interpretability. *arXiv preprint arXiv:1802.07810*, 2018.

[38] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144, 2016.

[39] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Anchors: High-precision model-agnostic explanations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

[40] Marko Robnik-Šikonja and Igor Kononenko. Explaining classifications for individual instances. *IEEE Transactions on Knowledge and Data Engineering*, 20(5):589–600, 2008.

[41] Jože M. Rožanec, Patrik Zajec, Klemen Kenda, Inna Novalija, Blaž Fortuna, and Dunja Mladenić. Xai-kg: knowledge graph to support xai and decision-making in manufacturing, 2021.

[42] Jože M. Rožanec and Dunja Mladenić. Semantic xai for contextualized demand forecasting explanations. *arXiv preprint arXiv:2104.00452*, 2021.

[43] Wojciech Samek and Klaus-Robert Müller. Towards explainable artificial intelligence. In *Explainable AI: interpreting, explaining and visualizing deep learning*, pages 5–22. Springer, 2019.

[44] A. Carlisle Scott, William J. Clancey, Randall Davis, and Edward H. Shortliffe. Explanation capabilities of production-based consultation systems. Technical report, STANFORD UNIV CA DEPT OF COMPUTER SCIENCE, 1977.

[45] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.

[46] Avanti Shrikumar, Peyton Greenside, Anna Shcherbina, and Anshul Kundaje. Not just a black box: Learning important features through propagating activation differences. *arXiv preprint arXiv:1605.01713*, 2016.

[47] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013.

[48] Dylan Slack, Sorelle A. Friedler, Carlos Scheidegger, and Chitradeep Dutta Roy. Assessing the local interpretability of machine learning models. *arXiv preprint arXiv:1902.03501*, 2019.

[49] Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda Viégas, and Martin Wattenberg. Smoothgrad: removing noise by adding noise. *arXiv preprint arXiv:1706.03825*, 2017.

[50] Kacper Sokol and Peter Flach. Limetree: Interactively customisable explanations based on local surrogate multi-output regression trees. *arXiv preprint arXiv:2005.01427*, 2020.

[51] Kacper Sokol, Alexander Hepburn, Raul Santos-Rodriguez, and Peter Flach. blimey: surrogate prediction explanations beyond lime. *arXiv preprint arXiv:1910.13016*, 2019.

[52] Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin Riedmiller. Striving for simplicity: The all convolutional net. *arXiv preprint arXiv:1412.6806*, 2014.

[53] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *International Conference on Machine Learning*, pages 3319–3328. PMLR, 2017.

[54] Jasper van der Waa, Marcel Robeer, Jurriaan van Diggelen, Matthieu Brinkhuis, and Mark Neerincx. Contrastive explanations with local foil trees. *arXiv preprint arXiv:1806.07470*, 2018.

[55] Feiyu Xu, Hans Uszkoreit, Yangzhou Du, Wei Fan, Dongyan Zhao, and Jun Zhu. Explainable ai: A brief survey on history, research areas, approaches and challenges. In *CCF international conference on natural language processing and Chinese computing*, pages 563–574. Springer, 2019.

[56] Petri Ylikoski and Jaakko Kuorikoski. Dissecting explanatory power. *Philosophical studies*, 148(2):201–219, 2010.

[57] Soyoung Yoo and Namwoo Kang. Explainable artificial intelligence for manufacturing cost estimation and machining feature visualization. *arXiv preprint arXiv:2010.14824*, 2020.

[58] Muhammad Rehman Zafar and Naimul Mefraz Khan. Dlime: A deterministic local interpretable model-agnostic explanations approach for computer-aided diagnosis systems. *arXiv preprint arXiv:1906.10263*, 2019.

[59] Patrik Zajec, Jože M Rožanec, Inna Novalija, Blaž Fortuna, Dunja Mladenić, and Klemen Kenda. Towards active learning based smart assistant for manufacturing. *arXiv preprint arXiv:2103.16177*, 2021.

[60] Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *European conference on computer vision*, pages 818–833. Springer, 2014.

[61] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2921–2929, 2016.

[62] Jianlong Zhou, Amir H. Gandomi, Fang Chen, and Andreas Holzinger. Evaluating the quality of machine learning explanations: A survey on methods and metrics. *Electronics*, 10(5):593, 2021.

[63] Luisa M. Zintgraf, Taco S. Cohen, Tameem Adel, and Max Welling. Visualizing deep neural network decisions: Prediction difference analysis. *arXiv preprint arXiv:1702.04595*, 2017.

DOI: 10.1561/9781680838770.ch6

Chapter 6

Confidence Assessment of AI Models in Simulated Industrial Environments

By Spyros Theodoropoulos, Dimitrios Dardanis, Georgios Sofianidis, Jože M. Rožanec, Panagiotis Tsanakas and Dimosthenis Kyriazis

The deployment of artificial intelligence (AI) solutions in simulated industrial environments, such as manufacturing production lines, minimizes the risks of physical damage caused by potential agent errors or malfunctions. Leveraging synthetic data generation and data augmentation techniques can increase the accuracy and robustness of an AI solution. To that end, artificially generated adversarial scenarios can be exploited to assess an AI agent's confidence level and quality. This chapter will present the state-of-the-art techniques that aim to increase the confidence assessment of manufacturing focused AI agents by spanning the fields of Reinforcement Learning, Explainable AI and Visual Analytics.

6.1 Introduction

At the heart of AI-powered Industry 4.0 systems lie modern machine learning (ML) methods such as Deep Learning and Reinforcement Learning. Such algorithms consume large amounts of data available through smart factory IoT sensors to support automated decision-making, process optimization and achieve improved working conditions for human workers.

However, these algorithms often exhibit complex stochastic behavior that can be difficult or impossible for humans to understand. On the other hand, manufacturing is a domain where the risk of a mistaken action can affect the well-being of the human worker, the integrity of the production process, or the quality of the end product. Therefore transparency, safety, and trustworthiness are fundamental properties that should be considered when designing AI models that interact physically with the shop floor.

Another important aspect is that as these models grow in complexity and sophistication, so does their need for larger sets of training data. These sets might be smaller than required in practice, suffer from poor quality, or misrepresent the actual real-life domain of the problem modeled.

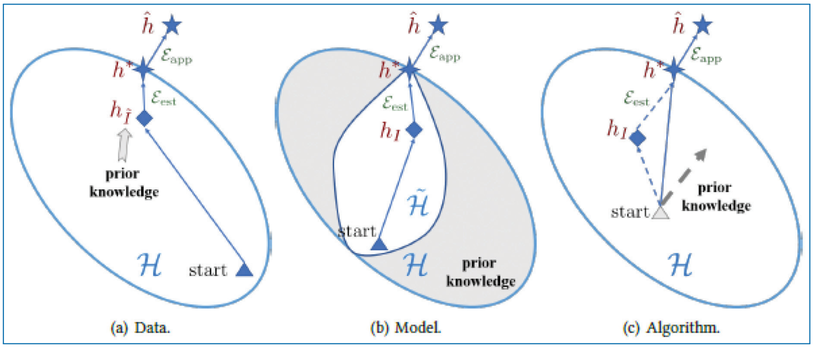
A simulation is a valuable tool for countering those issues, as it can augment the available input data, speed up the algorithm's training process, and help with its subsequent validation and robustification by producing original samples and scenarios. These can be used to enhance the algorithm's predictability and trustworthiness, especially when combined with confidence assessment methods and transparency enhancing techniques such as XAI and Visual Analytics. Their use in real-life AI deployments becomes necessary, as, despite their impressive results, algorithms based on Deep Learning and Reinforcement Learning still contain hidden aspects that humans cannot completely understand or control.

Confidence assessment techniques aim to make sure that an AI-powered machine will not act in a completely unpredictable way, endangering human workers or derailing the production process. We have seen XAI as a remedy providing us with understandable explanations of AI decisions. Another way is to try and mathematically quantify the confidence in an algorithm's decision or prediction. Such quantification can be achieved by modifying the algorithm to keep track of its confidence or by applying another AI algorithm or a statistical method over the outputs. The confidence assessment is a crucial step in larger model pipelines, as it determines whether a fallback method should be used or whether human intervention is necessary through human-in-the-loop and active learning techniques.

Our main focus in this chapter will center around well-studied manufacturing use-cases that are also predominant in the STAR H2020 project, namely defect detection and robotic pick-and-place tasks.

6.2 Simulated Reality

A Short Primer on Few-shot Supervised Learning



(a) Data. (b) Model. (c) Algorithm.

Figure 6.1. Categories of few-shot learning [1].

In the more straightforward case of supervised learning, the reality we are trying to simulate is the distribution generating the input images. Even though some classes of samples might be underrepresented, we can use the information we already have and transform or synthesize it to generate new samples. This process, namely data augmentation, is part of a ML sub-field called few-shot learning.

Few-shot learning (FSL) [1] is a set of techniques aimed at reducing the amount of training data needed for an algorithm and therefore tangent to Simulated Reality. There are three areas to address this problem: the input data, the model, and the optimization algorithm, as seen in Figure 6.1.

$\mathcal{H}$ is the hypothesis space or space of the family of models (e.g., all CNNs of a specific architecture). The optimization algorithm moves through this space by learning better and better parameters moving from the beginning to the learned hypothesis h_l (note that h_l depends on the training dataset), representing the final learned parameters. ε_{est} is the estimation error due to learning inefficiency (e.g., overfitting) and ε_{app} the approximation error, due to the limited capacity of the hypothesis space. What FSL is trying to do is bring the "start" point closer to h^* faster than some model training that would require an extensive samples collection. For example, model-based techniques such as transfer learning try to constrict $\mathcal{H}$ to $\tilde{\mathcal{H}}$, a smaller hypothesis space learned from another similar problem with a high chance of including h^*. On the other hand, the "algorithm" category tries to use prior knowledge over the learning rate and direction of the optimizer to decrease the number of model updates. With data augmentation, which is our main focus, we are trying to improve the accuracy gained by the model by synthesizing additional samples and bringing the final stage $h_{\tilde{l}}$ of the training closer to h^*.

6.2.1 Data Augmentation in Visual Defect Detection

In a use case such as visual quality inspection, a data augmentation approach is often necessary as defects rarely occur in manufacturing, and several classes of defects can be severely underrepresented. It is even more critical in the case when we want to check the robustness of an algorithm against previously unknown defect types or known occurrences that largely deviate from the training samples. There are various methods in the literature. The more traditional ones are based on image processing transformations. At the same time, more sophisticated methods attempt to create synthetic data using Variational Autoencoders, Generative Adversarial Networks (GANs) and Neural Style Transfer (NST).

The traditional approach to image data augmentation generates image data through various transformations, such as scaling, rotation, translation, shearing, blur, or illumination. However, those image-level transformations do not contribute sufficiently to the clearer separation between different classes, especially when the separation depends on higher level features [2]. To overcome the limitations of traditional image processing methods, Convolutional Variational Autoencoders (CVAEs) have been proposed. Those consist of two CNNs: the encoder, which maps the input image to a latent space of lower dimensionality, and the decoder, which generates a new varying reconstruction of the original image from the latent features. CVAEs have been used successfully in [3] to augment underpopulated defect classes on a dataset of metal surfaces. The classification output from a six-class CNN over the augmented dataset ended up having nearly perfect precision and recall scores, with a significant improvement over the pipeline without the CVAE.

Generative Adversarial Networks (GANs) are another important tool, especially for restoring balance in skewed datasets through synthetic image generation. They can efficiently address different kinds of imbalances such as inter-class, intra-class (e.g., person re-identification), object and pixel-level imbalances for segmentation tasks [4]. A concrete use-case for defect detection is presented in [5] using Wasserstein GANs. The method is used to detect burn-through and crack defects on welding joints which were found difficult to generate with traditional image processing methods. The data augmentation framework consists of two rival networks: a generator that generates fake images from random noise and a discriminator network tasked to distinguish between real and fake images. The original purpose of the generator network is to deceive the discriminator. However, here, it is also re-purposed to generate plausible, high-quality defect images. The generator consisted of six deconvolutional layers, with ReLU (and tanh for the last layer). The discriminator had four convolutional layers with leaky ReLU and produced up to seven defect classes. The synthetic data helped the network perform well with less than 10^4

original samples, producing misclassifications only between defect types and perfectly separating the "normal" class.

Neural Style Transfer attempts to fuse two images: the "style" image and the other "content" image. Starting from a random noise image, it tries to minimize the two losses for style and content simultaneously. While initial methods implemented only global style transfer, [6] uses the technique described in [7] to fuse defects only with local regions of the content image. The local fusion algorithm iterates over two steps, first using a patch-match method to find a patch in the style area similar to one in the content image area for replacement. Then, further training the network using histogram and variational loss functions improves smoothness on patch boundaries. The generated images were finally given as input to a segmentation network to detect defects in buttons and showed quite promising results against the vanilla segmentation and other generators based on CycleGAN [8] and histogram matching.

6.2.2 Simulated Reality in Reinforcement Learning for Robotic Control

In reinforcement learning (RL), simulation is of particular importance as a remedy for the sample complexity problem. RL algorithms, depending on the complexity of the domain, need many episodes of trial and error to learn efficient policies. These episodes can be expensive and risky to obtain, especially in high-risk real-world settings such as manufacturing sites, often making the use of simulation a necessity. Additionally, in an artificial simulation environment, forbiddingly risky policies can also be explored to help guarantee robustness or discover new, improved policies that would be inaccessible by sticking to a more conservative policy. Apart from sample complexity, simulation can also help with the transfer of knowledge. Learning in an abstracted environment that only retains the necessary elements for a particular task could make the learned policies more generalizable and adjustable to different settings (e.g., part handling of different parts or different production lines). Finally, simulation can provide an additional layer of safety where different possible consequences of an action can be observed and their results validated to avoid real-world accidents. Next, we will focus on approaches to bridge the sim-to-real gap mainly applied to the well-studied area of robotic grasping. This issue appears due to the accumulation of minor errors caused by the unavoidable inaccuracies in the simulation's physics model and visual rendering.

Domain adaptation is a set of techniques that help a learning model generalize to a target domain while trained with samples from different sources. In robotic grasping, simulation is the source domain, and the actual production line is the target. Domain adaptation is widely used in computer vision. It can be roughly distinguished into two categories: feature-level and pixel-level. Feature-level is usually based on adaptive feature extraction methods such as CNNs, which already have some degree of transferability between the simulation and reality domains. Including a domain-level similarity metric in their loss function, such as maximum mean discrepancy, can help enforce domain invariance when retraining in the new domain [9]. Pixel-level domain adaptation is mainly based on using GANs to restyle simulation images so that they look more similar to real ones [8]. Both of the above techniques can work well on Deep Reinforcement Learning algorithms that base their perception and action planning on CNNs. A good example is GraspGAN [10] which uses simulation with a hybrid adaptation method, combining Domain Adaptation Neural Networks (DANNs) with a novel batch-normalization technique. The proposed method achieved comparable or better performance to vanilla Deep RL with fifty times fewer real-world samples.

Domain randomization methods have also shown good results for the task of robotic grasping making simulation-only training feasible. The goal is to train the agent on a broader set of environmental conditions by introducing randomization in the simulated environment at training time. Given that the variability of the conditions is sufficient, the model trained in the simulation will be able to generalize in the real world. For instance, [11] uses randomization on the following types of features: addition of distracting objects of different shapes and sizes, object position and texture, the texture of background objects, camera position, orientation and field of view, number and position of lights and addition of different types of random noise. The trained model produced comparable results to real-world training, even though no real-world data was used.

A third approach is outlined in [12] where mapping is learned that can translate real-world images to their simulated equivalent. Domain randomization is utilized in this context to create the pairs of inputs and labels for the training of the translating network, called a Randomized-to-Canonical Adaptation Network (RCAN) [13]. The non-randomized simulated representation is referenced as the canonical representation and used to train the grasping algorithm. The RCAN is used during the real-world operation to translate real-world images into the canonical representation that the RL algorithm understands. A pipeline with RCAN and QT-Opt [14], a recent RL algorithm, manages to learn how to grasp previously unknown objects with high accuracy and a hundred times fewer episodes by training in the simulator only.

Finally, for the above methods to work, it is important to have an appropriate simulation environment. There are various open-source off-the-shelf options such as Gazebo [15], OpenSim [16], MuJoCo [17] and Bullet [18]. Choosing the right tool is very dependent on how much its features fit the requirements of the task at hand. A recent survey [19] seems to favor MuJoCo for the tasks of robotic grasping, though a model-based approach is used in the experiments. In GraspGAN [10] the robotic arm simulation was based on Bullet, giving priority on the amount of possible environmental diversity to ease the transfer to a real-world setting.

6.3 Confidence Assessment of AI Methods

The previous section illustrated how simulation can assist the ML training process and improve the performance of AI solutions. In real-world manufacturing environments, the notion of performance extends to the reliability and transparency of an AI system. This chapter will explore methods of confidence assessment as a way to improve an AI solution in a real-world industrial environment. To that end, the notion of confidence will be examined from two different perspectives. The first will examine methods of evaluating an AI algorithm's prediction confidence levels. In contrast, the second will focus on combining methodologies that enhance human cognition by providing deeper insights into AI's inner structures and decision-making processes.

6.3.1 Assessing the Confidence of Deep Neural Network Predictions

As we saw in the previous sections, DNNs and especially convolutional ones, are a powerful learning model. This has come at a cost, however, because as the model complexity of neural networks grows - which also brings an increase to their test accuracy - so does their overconfidence in their predictions [20]. In this section, we focus on the problem of classification. What we refer to as confidence in a classification setting is the maximal value of the last softmax layer, which determines the class of a given input. This is compared with the accuracy of the network for a given class. A good way to visualize this is reliability diagrams, where for different confidence ranges, the corresponding accuracies are shown [21]. Fig. 6.2 is an example for the five-layer LeNet model vs. the 110-layer ResNet on the CIFAR-100 dataset. The ideal would be for accuracy and confidence to be identical.

The main reason for this increasing miscalibration due to increasing model complexity is that DNNs also suffer from a more subtle case of overfitting. Namely, they tend to overfit the negative log-likelihood loss invisibly. In contrast, their visible generalization accuracy measured by a 0/1 loss seems to remain stable. This is a sign of unreliability that has limited DNN use in real-world safety-critical applications.

Many methods have been proposed to mitigate this overconfidence. The first category tries to adjust softmax outputs as a post-processing step to resemble the actual

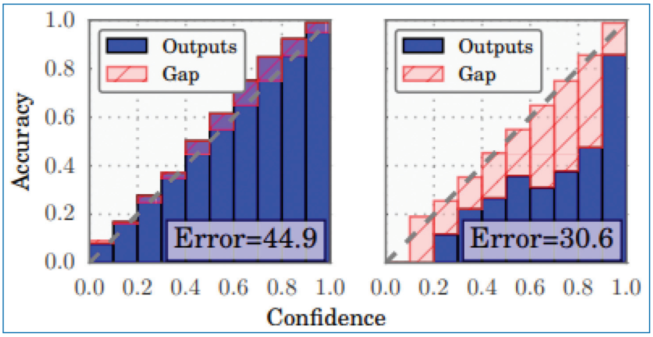


Figure 6.2. Reliability Diagram, LeNet vs. ResNet [21].

confidence probabilities (calibration methods) or follow an ordering where a higher value will correspond to higher "true" confidence. Histogram Binning [22], Isotonic Regression, and Bayesian Binning Quantiles (BBQ) [23] are example methods that solve optimization problems after the model training to bring softmax output close to their confidence values as estimated on a validation set. Platt Scaling [24] and its generalizations Matrix Weighting [20], and Temperature Scaling [25] are applied on the "logit" layer just before the softmax aiming to calibrate the weights of the final layer so that outputs are close to the validation set confidence probabilities. Temperature scaling is the most popular one. It does not influence the ordering of the class predictions guaranteeing the exact class prediction as before.

Alternatively, suppose one is interested only in an ordering of class confidence estimates. In that case, estimating the approximate distance from class boundaries [26] is another alternative. The second category of confidence assessment methods tries to make changes to the learning algorithm so that the training process is also constrained to output reasonable measures of the model's "true" confidence. Most notable is the addition of a term to the loss function that penalizes ordering inconsistencies in the output (pseudo-)probabilities [27]. Moreover, regularization techniques such as dropout, weight decay, label smoothing [28] and mixup [29] have been shown to improve confidence estimates.

6.3.2 Confidence and Reliability in Reinforcement Learning

Reinforcement learning applications in a real-world setting face significant challenges. The main difficulty lies in the fact that the agent acts autonomously. Therefore, any deviation from its usual way of operating will directly and potentially negatively impact the environment. An overview of possible Reinforcement Learning (RL) failures in practice was given by researchers of OpenAI in [30]. The most common pitfalls in the current state of the art are "negative side-effects" and "distribution shift". Negative side-effects occur when the agent tries to learn an objective function that focuses on a specific aim but ignores other important factors in the environment, usually considered as “common sense” knowledge by humans. “Distribution shift” refers to scenarios where the agent makes catastrophic mistakes while trying to adjust to changes in the underlying environment.

The easiest remedy for those problems in practical RL systems is to use a hard-coded reaction to potentially catastrophic events (e.g., forcing a robot to stop before colliding with an unexpected obstacle or keeping a drone high enough above the ground to avoid collision). However, as the domain complexity increases, it is increasingly challenging to incorporate all these scenarios beforehand. A few other ways of reducing exploration risks are described in [31], including filtering actions through reward thresholds and changing the total reward objective to include risk terms. An excellent example of the latter is also [32] which uses Conditional Value at Risk (CVaR) as a criterion for policy gradient optimization. Another alternative is to estimate lower confidence bound on the expected reward of a trajectory in an off-policy manner [33]. Finally, distribution shift can be countered by reachability analysis and the initial adoption of conservative policies, e.g., through robust policy improvement [34] or imitation learning [35].

Even if different improvements are added for making RL agents more risk-aware, it is still beneficial to extract confidence estimates from outside the agent, right between its choice for the best next action and the action’s execution in the real world. This can be achieved by anomaly detection techniques used for active learning [36, 37]. The inclusion of a human in the AI loop seems quite attractive for implementing real-life RL systems. However, such human intervention might be time-consuming or even impossible if the agent operates in tiny time scales. Human Intervention Reinforcement Learning (HIRL) [38] tries to solve this issue by complementing human intervention with AI. The human initially monitors the RL agent’s decisions. If one of them is catastrophic, it replaces it with a safe action and assigns it a large negative reward. A classifier, called a *Blocker*, monitors human intervention, learning to imitate the human’s blocking behavior. After a sufficiently long amount of time (and samples), the *Blocker* takes over the RL agent’s monitoring. This approach has shown to be very effective against catastrophic forgetting.[1] As the RL agent executes its policies for many episodes without taking any catastrophic actions, its value estimation becomes more and more optimistic, underestimating the negative value of those actions. The *Blocker*, trained from the human operator with a more constant perspective across time, can be there and remind the RL agent of forgotten pitfalls.

[1] Catastrophic forgetting describes the fact that an artificial neural network completely and abruptly forgets what was previously learned upon learning new information.

6.3.3 Visual Analytics and XAI Methods as a Way to "Peer Through the Black-box"

Despite the recent increase in the research of ML models, and especially DNNs and Graph Neural Networks (GNNs), their internal complexity is still often referred to as a "black box" [39, 40]. Due to their huge numbers of hyperparameters (up to thousands) combined with non-linear transformations, DNNs might seem obscure to important manufacturing stakeholders such as factory workers or production line managers, thus generating an issue of trust. That is why the notion of AI decision "confidence" as "the belief that a choice or a proposition is correct based on the available evidence" [41] is of paramount importance, especially in a dynamically changing environment, such as a production line. Additionally, the complementary notions of interpretability or explainability [42, 43] are key in fostering stakeholder understanding and trust. The field of explainable AI (XAI) aims to address this issue by generating valuable insights into the inner structures of ML algorithms and is a rapidly growing field of research that also includes visualization techniques to provide clarity to domain experts.

There is a vast amount of XAI methods described in the literature which can be classified based on different criteria such as (i) the complexity of interpretability, (ii) the scope of interpretability, and (iii) the level of dependency from the used AI model [44]. Complexity-related methods can be split into intrinsic explainability and post-hoc explainability. Intrinsic explainability is achieved by designing an AI model where the internal functioning is directly accessible to the user, making the model intrinsically interpretable (such as decision trees or linear regression). In contrast, post-hoc explainability accompanies the AI model by providing insights without knowing how the AI model works. Based on the scope of interpretability, global interpretability is referred to understanding the entire model behavior, and local interpretability is referred to understanding a single prediction. Finally, based on the level of dependency, post-hoc explanation methods can be split into model-agnostic that can be used to explain any kind of model and model-specific that are only suitable for specific model cases.

Especially for Deep Neural Networks (DNNs), several practical methods that fall under post-hoc explainability have been introduced to make those black-box models less opaque. [45] listed four families of explanation techniques: interpretable local surrogates, occlusion analysis, gradient-based techniques, and layerwise relevance propagation (LRP). Interpretable local surrogates aim to replace the decision function with a local surrogate model that is structured so that it is self-explanatory (e.g., the linear model). This approach is embodied in the LIME algorithm [46], which was successfully applied to DNN classifiers for images and text. Occlusion analysis is a perturbation-based technique where we repeatedly test the effect on the neural network output of occluding patches or individual features in the input image. As a result, a heatmap can be built highlighting locations where the occlusion has caused the most substantial function decrease. Gradient-based techniques involve gradient calculations in order to produce explanations. Integrated gradients [47] explain by integrating the gradient of the output along some trajectory in input space connecting some root point to the given data point. Another method is SmoothGrad [48] where the function's gradient is averaged over a large number of locations corresponding to small random perturbations of the original data point. Finally, the LRP method [49] makes explicit use of the layered structure of the neural network and operates in an iterative manner to produce the explanation. First, activations at each layer of the neural network are computed until we reach the output layer. The activation score in the output layer forms the prediction. Then, a reverse propagation pass is applied, where the output score is progressively redistributed, layer after layer until the input variables are reached. There are many applications of the above-listed methods in literature where XAI boosts the transparency and acceptance of high-performing black-box models. In the manufacturing domain, [50] described a post-hoc XAI analysis of deep learning for defect classification of Thin-film-transistor liquid-crystal display (TFT-LCD) panels. The authors used post-hoc techniques to produce heatmaps as explanations for a VGG network alongside decision trees and human interpretable rules. The results were successfully presented to domain experts in order to boost the acceptance of the AI model.

Working both synergistically and in parallel with XAI is the field of Visual Analytics (VA), which is associated with Information visualization and human interactions but also extends to other fields of Computer Science. Like [51] D. Keim *et al.* analyzed it is a field that combines Data Mining, ML, Data Management, and interactive visualizations in order to assist in decision-making processes for big heterogeneous data sets and provide useful insights to ML inner processes. The field of VA can be divided into two main areas: (i) *data analysis/data structuring*, which also includes Data Mining and ML techniques; (ii) *interactive visualizations*, which are responsible for data representation, explanations, and capturing complex human input in order to "translate" it into systems' actions. One of the very first who highlighted the importance of data visualization was [52] Card *et al.*, who defined information visualization as the use of computer-supported, interactive, visual representation of abstract data to amplify cognition. The notion of "meaning" in data representation systems is defined as the semantic information which can be extracted from the data representations and as semantics, the formalization which represents the meanings of the represented data. Nazemi *et al.* [53] extended Card's definition of semantics visualizations to computer-aided interactive representations for effective exploratory search, knowledge, domain understanding, and decision making. [51] Keim *et al.* noted that a VA solution needs to be expressive, effective, and appropriate; it is specified as [54] expressive if it represents exactly the information contained in the data (nothing more and nothing less); effective if it successfully represents the domain-specific and context-related information; and appropriate if it is beneficial in terms of cost/value ratio.

Traditional ML algorithms which are based on batch learning tend to suffer in large real-world environments. Expert information is not considered in the learning process, and therefore there is an issue of trust. [55] Wu *et al.* highlighted the importance of VA methods when applied in industrial environments since they constructively involve the human factor in decision-making processes providing more clarity of the ML decision. Furthermore, the statistical distribution for an industrial application (e.g., sensors) should be considered sensible to changes over time, as well as other unforeseen factors. A sensor could be re-calibrated, and therefore, its outputs will vary from those that an ML algorithm was trained with, in a previous time window. VA techniques that focus on building the ML model's input (processes before model build or after sudden statistical change), aim in assisting domain experts (e.g., factory workers, production line managers, etc.) to prepare their data and enhance human cognition in order to better comprehend the input to the ML algorithm. These techniques can be split in [56] two sub-categories: (i) interactively assessing the data quality in order to [57] generate ground truth labels from noisy-sourced labels; (ii) allowing domain experts to explore complicated Feature Spaces by adding another layer of semantic information to standard feature extraction techniques. Regarding interactive feature selection, the [58] Infuse framework suggests a powerful way of "correctly" identifying the most relevant features in large complicated feature spaces. VA solutions extend simple visualizations such as histograms of oriented gradients and allow users to interact with the underlying data. To that end, much research focuses on providing deeper insights to domain experts by providing *views* designed to highlight ML uncertainty or alternatively by establishing an iterative process for overcoming this issue and proposing solutions towards Active Learning methods. [59] Agocs *et al.* defined a "visual view" as the result of user interaction by inserting a query into a system that returns the corresponding subset of data. A view is the decomposition of directed and labeled graphs into multiple directed and weighted graphs of lesser dimensions. [60] Xu *et al.* provided a Visual Analytics approach to generate insight into the software structures of large-scale models; [61] Sibolla *et al.* presented a framework to analyze and visualize data streams from sensor observations; [62] KagNet proposes a textual inference framework to answer commonsense questions by exploiting knowledge graphs; [63] RetainVis proposed a method to increase user understanding of Recurrent Neural Networks and leverage domain expertise; [64] DeepEyes is a VA framework that assists the creation of DNNs by generating more insight on the different layers of the Neural Network and the filters that are triggered in each layer; [65] Legg *et al.*

proposed a VA methodology that incorporates human input in order to increase the confidence levels of an ML solution in dynamically changing environments.

6.4 Conclusion

Simulation-based methods combined with confidence assessment methods can provide an additional layer of safety over a deployed AI model. We saw that simulation can assist the training process and synthesize novel test cases to stress the model's current capabilities. Those test cases can be evaluated through the presented methods for supervised and reinforcement learning. They can then be used to inform model updates (e.g., sim2real) or be converted to a human interpretable form through XAI and VA to give deeper insight into the workings of the model.

A fully integrated Simulated Reality component in an Industry 4.0 environment is expected to support different families of learning algorithms, most importantly supervised (e.g., learning visual defects from a pre-labeled dataset of defects and non-defects) and reinforcement learning (e.g. autonomous robotic arm pick-and-place). Its core modules would be an algorithm-agnostic synthetic data generator coupled with a confidence assessment library, usable both during model training and real-world model testing. Besides outputting confidence scores corresponding to model predictions or actions, the confidence assessment component will also include the XAI and VA sub-components to assist technical and operational stakeholders in understanding the model's behavior. This will enable them to utilize their expertise in providing adjustments to improve the model's accuracy and general performance. Additionally, as described in the section about sim2real transfer in reinforcement learning, capabilities for automated model updates could be added trying to bridge the gap between simulation and reality and incorporate new knowledge acquired in simulation back to the real-world model.

Acknowledgements

This work has been carried out in the H2020 STAR project, which has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 956573.

References

[1] Y. Wang, Q. Yao, J. T. Kwok, and L. M. Ni, "Generalizing from a few examples: A survey on few-shot learning," *ACM Comput. Surv.*, vol. 53, June 2020.

[2] P. Pawara, E. Okafor, L. Schomaker, and M. Wiering, "Data augmentation for plant classification," 09 2017.

[3] J. P. Yun, W. C. Shin, G. Koo, M. S. Kim, C. Lee, and S. J. Lee, "Automated defect inspection system for metal surfaces based on deep learning and data augmentation," *Journal of Manufacturing Systems*, vol. 55, pp. 317–324, 2020.

[4] V. Sampath, I. Maurtua, J. J. A. Martín, and A. Gutierrez, "A survey on generative adversarial networks for imbalance problems in computer vision tasks," 2020.

[5] X. Le, J. Mei, H. Zhang, B. Zhou, and J. Xi, "A learning-based approach for surface defect detection using small image datasets," *Neurocomputing*, vol. 408, pp. 112–120, 2020.

[6] T. Wei, D. Cao, X. Jiang, C. Zheng, and L. Liu, "Defective samples simulation through neural style transfer for automatic surface defect segment," in *International Conference on Optical Instruments and Technology*, 2020.

[7] F. Luan, S. Paris, E. Shechtman, and K. Bala, "Deep painterly harmonization," *Computer Graphics Forum*, vol. 37, 2018.

[8] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 2242–2251, 2017.

[9] E. Tzeng, J. Hoffman, N. Zhang, K. Saenko, and T. Darrell, "Deep domain confusion: Maximizing for domain invariance," *ArXiv*, vol. abs/1412.3474, 2014.

[10] K. Bousmalis, A. Irpan, P. Wohlhart, Y. Bai, M. Kelcey, M. Kalakrishnan, L. Downs, J. Ibarz, P. P. Sampedro, K. Konolige, S. Levine, and V. Vanhoucke, "Using simulation and domain adaptation to improve efficiency of deep robotic grasping," 2018.

[11] J. Tobin, R. Fong, A. Ray, J. Schneider, W. Zaremba, and P. Abbeel, "Domain randomization for transferring deep neural networks from simulation to the real world," in *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 23–30, 2017.

[12] S. James, P. Wohlhart, M. Kalakrishnan, D. Kalashnikov, A. Irpan, J. Ibarz, S. Levine, R. Hadsell, and K. Bousmalis, "Sim-to-real via sim-to-sim: Data-efficient robotic grasping via randomized-to-canonical adaptation networks," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12619–12629, 2019.

[13] S. James, P. Wohlhart, M. Kalakrishnan, D. Kalashnikov, A. Irpan, J. Ibarz, S. Levine, R. Hadsell, and K. Bousmalis, "Sim-to-real via sim-to-sim: Data-efficient robotic grasping via randomized-to-canonical adaptation networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12627–12637, 2019.

[14] D. Kalashnikov, A. Irpan, P. Pastor, J. Ibarz, A. Herzog, E. Jang, D. Quillen, E. Holly, M. Kalakrishnan, V. Vanhoucke, *et al.*, "Qt-opt: Scalable deep reinforcement learning for vision-based robotic manipulation," *arXiv preprint arXiv:1806.10293*, 2018.

[15] N. Koenig and A. Howard, "Design and use paradigms for gazebo, an open-source multi-robot simulator," in *2004 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) (IEEE Cat. No.04CH37566)*, vol. 3, pp. 2149–2154, 2004.

[16] S. L. Delp, F. C. Anderson, A. S. Arnold, P. Loan, A. Habib, C. T. John, E. Guendelman, and D. G. Thelen, "Opensim: Open-source software to create and analyze dynamic simulations of movement," *IEEE Transactions on Biomedical Engineering*, vol. 54, no. 11, pp. 1940–1950, 2007.

[17] E. Todorov, T. Erez, and Y. Tassa, "Mujoco: A physics engine for model-based control," *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 5026–5033, 2012.

[18] E. Coumans, "Bullet physics simulation," p. 1, 07 2015.

[19] T. Erez, Y. Tassa, and E. Todorov, "Simulation tools for model-based robotics: Comparison of bullet, havok, mujoco, ode and physx," in *2015 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 4397–4404, 2015.

[20] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," in *Proceedings of the 34th International Conference on Machine Learning – Volume 70*, ICML'17, p. 1321–1330, JMLR.org, 2017.

[21] A. Niculescu-Mizil and R. Caruana, "Predicting good probabilities with supervised learning," in *Proceedings of the 22nd International Conference on Machine Learning*, ICML '05, (New York, NY, USA), p. 625–632, Association for Computing Machinery, 2005.

[22] B. Zadrozny and C. Elkan, "Transforming classifier scores into accurate multiclass probability estimates," in *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '02, (New York, NY, USA), p. 694–699, Association for Computing Machinery, 2002.

[23] M. P. Naeini, G. F. Cooper, and M. Hauskrecht, "Obtaining well calibrated probabilities using bayesian binning," in *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, AAAI'15, p. 2901–2907, AAAI Press, 2015.

[24] J. Platt, "Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods," *Adv. Large Margin Classif.*, vol. 10, 06 2000.

[25] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," in *NIPS Deep Learning and Representation Learning Workshop*, 2015.

[26] G. F. Elsayed, D. Krishnan, H. Mobahi, K. Regan, and S. Bengio, "Large margin deep networks for classification," 2018.

[27] J. Moon, J. hyo Kim, Y. Shin, and S. Hwang, "Confidence-aware learning for deep neural networks," in *ICML*, 2020.

[28] R. Müller, S. Kornblith, and G. E. Hinton, "When does label smoothing help?," in *NeurIPS*, 2019.

[29] S. Thulasidasan, G. Chennupati, J. Bilmes, T. Bhattacharya, and S. Michalak, "On mixup training: Improved calibration and predictive uncertainty for deep neural networks," *ArXiv*, vol. abs/1905.11001, 2019.

[30] D. Amodei, C. Olah, J. Steinhardt, P. F. Christiano, J. Schulman, and D. Mané, "Concrete problems in ai safety," *ArXiv*, vol. abs/1606.06565, 2016.

[31] M. Pecka and T. Svoboda, "Safe exploration techniques for reinforcement learning – an overview," in *MESAS*, 2014.

[32] A. Tamar, Y. Glassner, and S. Mannor, "Policy gradients beyond expectations: Conditional value-at-risk," *ArXiv*, vol. abs/1404.3862, 2014.

[33] P. S. Thomas, G. Theocharous, and M. Ghavamzadeh, "High-confidence off-policy evaluation," in *AAAI*, 2015.

[34] J. F. Fisac, N. F. Lugovoy, V. R. Royo, S. Ghosh, and C. Tomlin, "Bridging hamilton-jacobi safety analysis and reinforcement learning," *2019 International Conference on Robotics and Automation (ICRA)*, pp. 8550–8556, 2019.

[35] C. Tessler, Y. Efroni, and S. Mannor, "Action robust reinforcement learning and applications in continuous control," in *ICML*, 2019.

[36] D. Krueger, J. Leike, O. Evans, and J. Salvatier, "Active reinforcement learning: Observing rewards at a cost," *ArXiv*, vol. abs/2011.06709, 2020.

[37] P. F. Christiano, J. Leike, T. Brown, M. Martic, S. Legg, and D. Amodei, "Deep reinforcement learning from human preferences," in *NIPS*, 2017.

[38] W. Saunders, G. Sastry, A. Stuhlmüller, and O. Evans, "Trial without error: Towards safe reinforcement learning via human intervention," in *AAMAS*, 2018.

[39] A. Weller, "Challenges for transparency," 07 2017.

[40] T. Zahavy, N. Ben-Zrihem, and S. Mannor, "Graying the black box: Understanding dqns," in *Proceedings of The 33rd International Conference on Machine Learning* (M. F. Balcan and K. Q. Weinberger, eds.), vol. 48 of *Proceedings of Machine Learning Research*, (New York, New York, USA), pp. 1899–1908, PMLR, 20–22 Jun 2016.

[41] A. Pouget, J. Drugowitsch, and A. Kepecs, "Confidence and certainty: distinct probabilistic quantities for different goals," *Nature neuroscience*, vol. 19, no. 3, pp. 366–374, 2016.

[42] G. Montavon, W. Samek, and K.-R. Müller, "Methods for interpreting and understanding deep neural networks," *Digital Signal Processing*, vol. 73, pp. 1–15, 2018.

[43] T. Miller, "Explanation in artificial intelligence: Insights from the social sciences," *Artif. Intell.*, vol. 267, pp. 1–38, 2019.

[44] A. Adadi and M. Berrada, "Peeking inside the black-box: a survey on explainable artificial intelligence (xai)," *IEEE access*, vol. 6, pp. 52138–52160, 2018.

[45] W. Samek, G. Montavon, S. Lapuschkin, C. J. Anders, and K.-R. Müller, "Explaining deep neural networks and beyond: A review of methods and applications," *Proceedings of the IEEE*, vol. 109, no. 3, pp. 247–278, 2021.

[46] M. T. Ribeiro, S. Singh, and C. Guestrin, ""why should i trust you?" explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 1135–1144, 2016.

[47] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic attribution for deep networks," in *International Conference on Machine Learning*, pp. 3319–3328, PMLR, 2017.

[48] D. Smilkov, N. Thorat, B. Kim, F. Viégas, and M. Wattenberg, "Smoothgrad: removing noise by adding noise," *arXiv preprint arXiv:1706.03825*, 2017.

[49] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, and W. Samek, "On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation," *PloS one*, vol. 10, no. 7, p. e0130140, 2015.

[50] M. Lee, J. Jeon, and H. Lee, "Explainable ai for domain experts: a post hoc analysis of deep learning for defect classification of tft–lcd panels," *Journal of Intelligent Manufacturing*, pp. 1–13, 2021.

[51] D. Keim, G. Andrienko, J.-D. Fekete, C. Görg, J. Kohlhammer, and G. Melançon, *Visual Analytics: Definition, Process, and Challenges*, pp. 154–175. Berlin, Heidelberg: Springer Berlin Heidelberg, 2008.

[52] S. K. Card, J. D. Mackinlay, and B. Shneiderman, eds., *Readings in Information Visualization: Using Vision to Think*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1999.

[53] K. Nazemi, D. Burkhardt, E. Ginters, and J. Kohlhammer, "Semantics visualization – definition, approaches and challenges," *Procedia Computer Science*, vol. 75, pp. 75–83, 2015. 2015 International Conference Virtual and Augmented Reality in Education.

[54] W. Cui, "Visual analytics: A comprehensive overview," *IEEE Access*, vol. 7, pp. 81555–81573, 2019.

[55] W. Wu, Y. Zheng, K. Chen, X. Wang, and N. Cao, "A visual analytics approach for equipment condition monitoring in smart factories of process industry," in *IEEE Pacific Visualization Symposium, PacificVis 2018, Kobe, Japan, April 10–13, 2018*, pp. 140–149, IEEE Computer Society, 2018.

[56] Y. Jun, Y. W. Chen Changjian, X. J. Liu Mengchen, and L. Shixia, "A survey of visual analytics techniques for machine learning," *Computational Visual Media*, vol. 7, 2021.

[57] N. Q. V. Hung, D. C. Thang, M. Weidlich, and K. Aberer, "Minimizing efforts in validating crowd answers," in *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data*, SIGMOD '15, (New York, NY, USA), p. 999–1014, Association for Computing Machinery, 2015.

[58] J. Krause, A. Perer, and E. Bertini, "INFUSE: interactive feature selection for predictive modeling of high dimensional data," *Visualization and Computer Graphics, IEEE Transactions on*, vol. 20, pp. 1614–1623, Dec 2014.

[59] A. Agocs, D. Dardanis, J. L. Goff, and D. Proios, "Interactive graph query language for multidimensional data in collaboration spotting visual analytics framework," *CoRR*, vol. abs/1712.04202, 2017.

[60] Y. Xu, D. Wang, T. Janjusic, W. Wu, Y. Pei, and Z. Yao, "A web-based visual analytic framework for understanding large-scale environmental models: A use case for the community land model," in *ICCS*, 2017.

[61] B. H. Sibolla, S. Coetzee, and T. L. Van Zyl, "A framework for visual analytics of spatio-temporal sensor observations from data streams," *ISPRS International Journal of Geo-Information*, vol. 7, no. 12, 2018.

[62] B. Y. Lin, X. Chen, J. Chen, and X. Ren, "KagNet: Knowledge-aware graph networks for commonsense reasoning," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, (Hong Kong, China), pp. 2829–2839, Association for Computational Linguistics, Nov. 2019.

[63] B. C. Kwon, M.-J. Choi, J. T. Kim, E. Choi, Y. B. Kim, S. Kwon, J. Sun, and J. Choo, "Retainvis: Visual analytics with interpretable and interactive recurrent neural networks on electronic medical records," *IEEE transactions on visualization and computer graphics*, vol. 25, no. 1, pp. 299–309, 2019.

[64] N. Pezzotti, T. Höllt, J. Van Gemert, B. P. Lelieveldt, E. Eisemann, and A. Vilanova, "Deepeyes: Progressive visual analytics for designing deep neural networks," *IEEE Transactions on Visualization and Computer Graphics*, vol. 24, no. 1, pp. 98–108, 2018.

[65] L. Phil, J. Smith, and A. Downing, "Visual analytics for collaborative human-machine confidence in human-centric active learning tasks," *Human-centric Computing and Information Sciences*, vol. 9, 2019.

DOI: 10.1561/9781680838770.ch7

Chapter 7

The Human-Digital Twin in the Manufacturing Industry: Current Perspectives and a Glimpse of Future

By Elias Montini, Niko Bonomi, Fabio Daniele, Andrea Bettoni, Paolo Pedrazzoli, Emanuele Carpanzano and Paolo Rocco

The need to comply with shorter product life-cycles, diversified market demands and increased global competitiveness is leading to a dramatic increase in production systems requirements in terms of flexibility and responsiveness. Industry 4.0 and its push for digitalisation are becoming a pervasive reality impacting almost each phase of the company's life cycle, from business strategy and process design to daily operational activities. As a resulting drawback, an increasing burden is put on the workers, who are requested to operate and interact with complex systems, under challenging conditions. Furthermore, decision-making and control systems consider humans as an external and unpredictable element. New technological solutions are arising to address such challenges. Digital Twins can be applied to represent humans in the digital world, including their intents, behaviours, conditions and emotions, providing the ground for human-aware operations and planning. This chapter aims to provide an overview about most recent advancements and results applying digital twins to support the design, the implementation and the operations of human-centric production systems.

7.1 Introduction

From its origin, the manufacturing industry followed a continuous evolution that allowed to reach an unprecedented level of performance to satisfy increasingly demanding customers. However, despite the steep technological evolution, humans are still the fundamental resource of any production system.

In the manufacturing industry, more than 70% of tasks are still done manually, thus making humans accountable for generating most of the value [1]. Moreover, it is also expected that, by 2025, the average time spent by humans and machines at work will be the same as today [2]. Therefore, the human factor is and will be an essential dimension during design, deployment and operation of manufacturing systems. This notwithstanding, in the era of Industry 4.0, where Cyber-Physical Systems (CPSs) rule the roost, most systems still consider the human as an external and almost unpredictable element while human intents, physical states, characteristics and actions should be integrated into their design and operation [3]. The "human factor" has been often considered, with reference to the design of better ergonomics, so as to optimise interactions between workers, their tasks and the physical elements of the factory. However, this approach shows different limitations. First, it is performed offline, and it does not allow to dynamically adapt the production system to the current situation. Second, it considers human as an immutable entity, whose behaviours and conditions do not evolve in time. Finally, it takes care only of physical interactions between human and machines, without considering more complex relations, and the impacts that such relations may have one on each other and on the production system as a whole.

In recent years, the awareness that this approach demonstrates several shortcomings has come to the surface. In digital representations of the factories, where more and more complete and realistic twins of machines and equipment are included, the human has been almost neglected. Nevertheless, human's characteristics, behaviours and psychophysical conditions have a relevant impact on the performance and operations of a production system. Neglecting these elements in the digital representation of the production environment creates many limitations, especially considering the high emphasis that the concepts of Industry 5.0 [4] and Operator 4.0 brings on human factors. Therefore, in order to take the digital representation of the production systems a step further, humans have to be modelled and included. To realise this ambitious goal, the creation of a Human Digital Twin (HDT) is a relevant challenge.

In this regard, this work aims to explore recent advancements in the digital representation of humans in production systems. The Chapter is structured as follows. Section 7.2 introduces the concept of Operator 4.0 and of Human Cyber-Physical-Systems (H-CPSs). Section 7.3 provides a review of its main applications in the manufacturing industry. Section 7.4 highlights the most relevant technologies to realise HDT that can be applied in the manufacturing sector. Section 7.5 highlights the approach adopted in the STAR project towards the HDT development and adoption. Finally, concluding remarks are highlighted in Section 7.6.

7.2 The Operator 4.0 and Human Cyber-Physical-Systems

Each industrial revolution has led to a deep and significant transformation in the way production systems, engineering processes and manufacturing products are designed. Unavoidably, the operators' activities, responsibilities and nature of work have also undergone drastic modifications.

The spreading of the approaches and technologies belonging to the Industry 4.0 paradigm has radically changed the operators' roles in the modern factory environment leading to the definition of the Operator 4.0, known as "*a smart and skilled operator who does not only perform cooperative work with robots but also works aided by machines as and if needed by means of Human-Cyber-Physical-Systems, advanced human-machine interaction technologies and adaptive automation towards achieving human-automation symbiosis work systems*" [5]. In this sense, the Operator 4.0 is considered as a hybrid resource coming from the relationship between the human and machines, where the focus is, on the one hand, treating automation as a further extension of the human's physical, interaction and cognitive capabilities and, on the other hand, considering human as a precious source of information within the smart production environment [6].

The Operator 4.0 vision involves an operator surrounded by a digital work system which is perfectly suited for workers with different skills, capabilities, preferences, and background and also capable of maximizing their motivation and performance [7]. Specifically, the Operator 4.0 vision is strictly correlated with the concept of Human-Cyber-Physical Systems intentionally designed to enable the collaboration between humans and machines. A H-CPS is a system engineered to: (a) improve human abilities to dynamically interact with machines in the cyber and physical worlds employing intelligent human-machine interfaces, using human-computer interaction techniques designed to fit the operators' cognitive and physical needs, and (b) improve human physical, sensing and cognitive capabilities, using various enriched and enhanced technologies (e.g., wearable devices) [8].

The current conception of H-CPS is the result of a long path marked by the technological findings of industrial evolutions. When the manufacturing sector entered the era of digitalization in the end of the 20th century, the digital world interposed and started to link together human and physical system giving life to the H-CPS. The latter considerably enhanced computation, efficiency, control, precision and capabilities of production systems allowing to connect the digital, the physical, and the human side of these systems [9]. In the last decades, huge progress has been made in information technologies leading to the breakthrough of the Digital Twins (DTs), adding further features to the H-CPS.

The National American Space Agency (NASA) introduced the DT as "an integrated multi-physics, multi-scale, probabilistic simulation of a flying vehicle or systems" [10]. From that moment, the concept has been applied in many domains, including manufacturing. With this wide adoption, the concept evolved. Nowadays, observing the most recent definitions, and also the scope of this work, the DT can be defined as *"A digital twin is a digital replica of a physical entity. The Digital twin refers to actual or potential status of physical assets, processes, people, systems and devices that can be used for various purposes: planning, optimization, what-if analysis, monitoring … The DT is a dynamic virtual representation of a physical object/system across its lifecycle, using real-time data to enable understanding, learning and reasoning"*. It is necessary to emphasize that the DT should not to be confused with a simulation model. The DT has to be a high-fidelity virtual replica of a physical entity with real-time two-way communication supporting simulation and decision-making for product service enhancement [11].

Throughout the H-CPS evolution phases, the digital system has experimented diverse major enhancements meanwhile its connection with the human has remained almost unchanged. Despite the hundred examples of DTs applied to manufacturing products, devices, and machines, only few works addressing the HDT can be found even though the operator continues to be a key resource with relevant impacts on the performance of the manufacturing system. For these reasons, to bring the H-CPS a step further, humans have to be modelled and included in the digital world, together with the existent DTs.

7.3 The Human Digital Twin: A Review of Existing Applications

According to Segan and colleagues, the HDT can include models fed by dynamic and real-time data merged with static or quasi-static ones, enabling a comprehensive representation of the human entity [12]. To gather real-time information, activities and behaviours performed by humans have to be recorded. Moreover, these real-time data have to be compared with historical data, which are stored and formalised together with information describing human characteristics and conditions [13]. The HDT has not only to provide a digital representation of anthropometric and physiological features but also a representation of a person's inner state [14].

In [15] a meta-model is defined to realise a modular and tailored HDT comprehensive of all the entities that need to be modelled to create a HDT. These include worker's characteristics, medical, emotional and psychophysical conditions, psychophysical and geospatial parameters, contextual, functional and decision models. The HDT is a necessary technology to facilitate human worker integration in an Industry 4.0 environment to address communication, data aggregation, simulation and scheduling [16]. The emerging applications of HDT realized in the manufacturing industry in the last 5 years mainly include workers monitoring, production planning and scheduling, human-robot collaboration and adaptive automation.

Worker Well-being Monitoring

Thanks to the miniaturization and reduction of costs, the adoption of wearables and sensors, has been growing also in the industrial context to investigate workers' conditions and well-being. Employee's well-being is a key factor in determining the organization's long-term competitiveness and it is also directly related to production efficiency. The cumulative effect of positive impacts on the human factor brings economic benefit through productivity increase, scrap reduction and decrease of absenteeism. Few research works have been recently developed, where workers' physiological data are used to infer the insurgence of phenomena such as fatigue [17, 18] and mental stress [19] which, in turn, have a relevant impact on process performance. Another research line adopted eye-trackers, together with wearables and cameras to estimate worker's attention and stress levels, to understand assembly sequence and to identify the criticalities in the product design affecting the assembly process [20].

Production Planning and Allocation

In labour-intensive production systems, workers allocation is one of the key activities as capacity and worker skills are the main factors that affect the production rate [21]. The exclusion of such elements from the problem definition may result in poor performance. To this end, the HDT has been applied to support production planning and allocation. [22, 23] propose a HDT to store, organize and communicate workers' skills, preferences, virtualized personality and to enable humans to take part to a decentralized computational decision-making process which leads to an improved task scheduling. However, in this context and application, it is fundamental to consider that, despite the different taxonomies and methods that exist to assess workers skills like ESCO and O*NET, humans learn and grow, being able to deal with novel challenges. This is a relevant challenge to consider in realising a HDT supporting production planning and scheduling.

Human-Robot Collaboration and Adaptive Automation

In the last decades, adaptive controls systems have been explored, including workers features within real-time control loops [24–26]. The HDT is crucial for such kind of approaches. [27] used a Microsoft Kinect sensor to identify and track a worker within a work cell to identify the possible collisions areas with a robot. This information has been used to optimize in real-time robot trajectories in order to reduce collisions. In [28], a HDT, has been adopted to monitor worker fatigue and mental stress by introducing a physiological monitoring system and a smart decision-maker to adjust the level of support offered through a collaborative robot (cobot).

Ergonomics Analysis and Layout Design

Many examples exist where the HDT has been used to analyse the work cell ergonomics and to define the best layout considering worker characteristics, anthropometric ones above all. Many examples exist applying off-line simulation of humans [29]. Moreover, a few real-time examples can be found, where worker posture and characteristics are computed using motion capture tools [30] and even used to feed control systems to perform collaborative and ergonomic tasks [31].

Other Context Applications Outside the Manufacturing Industry

Extending the scope of the use of HDT outside the manufacturing industry, it is possible to find interesting applications. A digital representation of construction workers has been created collecting automatically physiological parameters (e.g., human heart rate, upper body posture angle, traveling speed) to identify workload severity [32]. In the same sector, workers' thoracic posture and spatio-temporal data have been used to estimate activity types and assess productivity in real-time [33]. In the medical and fitness fields, some applications implement the automatic exchange of data relying on wearable devices to compute health conditions [34] and predict athletes' performance [35].

7.4 The Technological Framework for Human Digital Twins

Taking inspiration from works describing the technologies and architectures behind the creation of a HDT [28, 36] and, more in general from those dedicated to DT in a broader sense [37, 38], 3 technological layers, as shown in Figure 7.1, are fundamental to realise a HDT, embracing all the cutting edge technologies end methodologies that enable the digitization of a worker, together with his/her characteristics, conditions and behaviours.

The **Sensors Layer** represents a connection from the physical to the digital world, in charge of the **creation of data** from the physical production system and of the **communication** in quasi-real time using standard data formats. In the case of the human, the installed hardware to realise part of such connection are mainly wearable sensors that fetch psychophysical parameters like Hearth Rate (HR), Skin Conductance (SC) or Galvanic Skin Response (GSR), while the machinery present at the shop floor needs to be fitted with sensors that generate useful data to be shared with the upper layers. In the case of a HDT, to connect wearables and sensors with the upper layers in many cases a gateway is the best solution to be adopted [39]. This is particularly helpful in case of limited battery capacity on wearables side, allowing to have these devices smaller, more comfortable and cheaper. A similar concept to the gateway is applied to machinery and robots, since most of them use industrial standard protocols like OPC-UA, Euromap or "simply" offer a PLC communication interface. In this case, the gateway acts as a middleware that translates the machine protocols into the unique standard communication protocol known by the whole HDT. This approach allows to realise a simple plug and play architecture, requiring only to build specific gateways without changing anything upstream. Gateways can be installed on many types of devices, including smartphones, computers or a simple raspberry. The key factor to consider is that this device needs

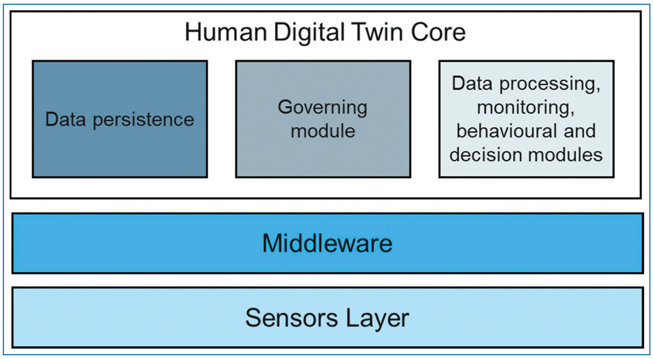


Figure 7.1. Abstract digital twin architecture.

to be able to join a network connection and to have enough capacity to keep a constant flow of data to and from the Middleware, which is the next layer that resides between the HDT Core and the Sensors Layer. In this regard, the NTN 5G is a relevant enabler that allows to have low latency and high reliability data streams, without the need of wired connections [40]. **Low-level computation** or **pre-processing** of data may resides in the Sensing Layer, depending on computational requirements and applications. It is necessary to consider that, if not properly performed, pre-processing may lead to information loss. However, if executed correctly, relevant benefits can be obtained in terms of communication efficiency, security and scalability.

The **Middleware** enables an active connection from and to the HDT Core. Since, it has to manage withstand high-frequency data coming from multiple sensors installed in the Sensors Layer, the Middleware has to be well designed and implemented to **exploit data flows** from physical to digital layer. It has to be capable to empower the **coherent integration** between the models and the physical architecture, enabling a seamless usage of data for **verification** and **validation** of complex behaviours. There are various tools and technologies that support the implementation of this layer. The most commonly used solutions are based on MQTT [41], a well-known and established data exchange protocol already widely used in the IOT and industrial world. One of the alternatives that implement MQTT is Apache Kafka [42] which offers many useful functions including short-term memory to keep a backlog of the last exchanged messages. However, in some use cases, a more simple and light approach is suggested to favour of performance, such as the one adopted by Mosquitto [43].

The third layer is the core of the HDT, where **data persistence**, **governing module** and the various **data processing**, **monitoring**, **behavioural** and **decision modules** are located. In the data persistence, it is possible to find all the data **storage functionalities** including databases that contain the data-models, describing the entities and features, and historical storage for the data coming from the sensors. For this latter functionality, it is recommended to use a time-series database like InfluxDB [44] or QuantumLeap [45], optimized for this kind of data. Finally, there are functional modules, which are capable to **process**, **simulate**, **predict**, **reason** and **decide**. The best approach to include data processing, monitoring, behavioural and decision modules is to adopt a series of plug and play services. The key benefit of such approach, obtained thanks to a proper architecture design, is that it can easily be removed, added or extended [46]. An alternative is the development of specific plugins. However, this requires the development of dedicated SDK and imposes a specific programming language. Meanwhile using services, the developer is free to select whichever programming language since the communication interface is universal. Modules can be of different nature. They can be focused on the post-processing of the data that transit on the Middleware. This can include data validation, classification, aggregation, sorting, and cleaning. Monitoring models focus on specific features and attributes, monitoring their evolution and elaborating raw data to compute more complex information and detecting possible deviations. The decision modules identify decisions through the HDT in order to intervene in the digital and/or in the physical world. The behavioural modules elaborate the current status of the HDT to make predictions and simulate its evolution.

Artificial Intelligence (AI) plays a relevant and prominent role in the creation of such kind of modules. One of the main goals of AI is to build software systems, models and algorithm capable to perform complex tasks [47]. The behaviours and variables that involve a human are much more complex than those that can characterize machines and robots. Furthermore, humans in a system are infinitely more unpredictable and have greater degrees of freedom than a machine, which is usually stationary, always performs the same tasks, has a set of standard components. For these reasons, AI is of major relevance for creating a HDT, which allows to create models without actually knowing the relationship between the inputs and the outputs. Without the help of AI, it would be very complex to define heuristic relationships between, for example, given physiological data such as HR, HRV, etc. and physical or mental stress. However, AI alone is not enough. It is necessary to consider that AI needs valuable data to build effective models and algorithms. Moreover, when humans are modelled with AI, it is also fundamental to consider related ethics and explainability issues.

7.5 The STAR Approach Towards Human Digital Twins for Manufacturing Applications

To realise a step ahead in the sustainability of production systems it is necessary to evolve current approaches and technologies mainly oriented to digitalisation aspects by including also human factors. Therefore, in the STAR project, a novel approach to the digitalisation of humans is proposed. To get the most from workers and the production systems, where a multitude of other players act, including robots, automation systems, AGVs and machines, it is necessary to include a more sophisticated and complete representation of workers. Humans and machines have not to be considered independent, but as collaborative entities that complement their capacities in order to achieve improved manufacturing performance, and their historical data, status and evolution must be available for analysis and optimization. To achieve such a goal, the digital representation of workers proposed by STAR allows to include contextual data (e.g., assigned job, current workplace, current shift, training program), quasi-static data (e.g., worker needs, skills, height, age), real-time sensor data (e.g., heart rate, heart rate variability, galvanic skin response, temperature) and dynamic data (e.g., fatigue level, emotions, position, current activity). Thanks to distributed sensing solutions that gather data about the workers and the surrounding workplace, it is possible to build a digital replica of a human. The STAR's HDT includes: (i) sensing modules to easily connect sensors and gather data from the shop-floor; (ii) a specific component to collect, structure, store and use data to realise the digital representation; (iii) a set of AI and non-AI modules to elaborate sensors data and compute complex features.

The STAR's HDT can be considered as a single source of truth of workers-related data. It offers a centralized access point to exploit wider set of workers' related data. STAR creates a digital representation of the workers, seamlessly integrated with production system DTs, that can be exploited by AI-based modules to compute complex features, feeding and enriching the HDT itself, or to make better decisions, dynamically adapting automation systems behaviour targeting both production performance and workers' safety and well-being. The STAR's HDT is organised in 3 main technological layers, according to the architecture introduced in Section 7.4 and is composed of the following components, as depicted in Figure 7.2.

- **Shop-floor entities, agents and gateways (Sensors Layer):** sensors, wearables and PLCs collect and stream data from the shop-floor. To facilitate data gathering from the workers and the production system entities, the HDT integrates agents and gateways to ensure the data collection, harmonisation

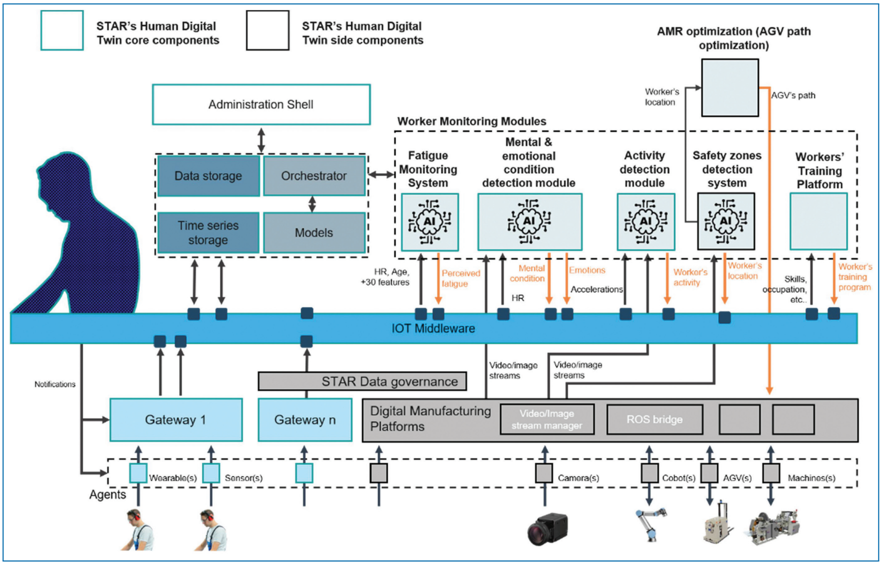

STAR's Human Digital Twin core components
STAR's Human Digital Twin side components
AMR optimization (AGV path optimization)
Worker's location
AGV's path
Administration Shell
Worker Monitoring Modules
Data storage
Orchestrator
Time series storage
Models
Fatigue Monitoring System
Mental & emotional condition detection module
Activity detection module
Safety zones detection system
Workers' Training Platform
AI
HR, Age, +30 features
Perceived fatigue
Mental condition
HR
Emotions
Accelerations
Worker's activity
Worker's location
Skills, occupation, etc..
Worker's training program
IOT Middleware
Video/image streams
Video/image streams
Notifications
STAR Data governance
Gateway 1
Gateway n
Digital Manufacturing Platforms
Video/Image stream manager
ROS bridge
Wearable(s)
Sensor(s)
Camera(s)
Cobot(s)
AGV(s)
Machines(s)
Agents

Figure 7.2. STAR's HDT architecture.

and accessibility from heterogeneous sources and to create bridges between these sources and the upper layers.

- **IoT Middleware (Middleware):** this layer supports M2M connection and it is based on the MQTT lightweight messaging protocol. It allows bi-directional communication under a publish-subscribe mechanism and the organisation of important amounts of heterogeneous data into multiple topics. Each user has a set of channels where data are streamed to and accessed by the modules that need them for further computations.
- **Data storage and Time Series Data Storage (Human Digital Twin Core – Data Persistence):** in the data storage all the structure and core information about the HDT are stored. In addition, the workers' quasi-static data are persisted in this component. Meanwhile, the Time Series Data Storage acts as a backlog of sensors data, in which the various entities of the HDT can access in order to make predictions or extract feature for computations.
- **Orchestrator and Models (Human Digital Twin Core – Governance module):** this component is responsible to manage all the entities in the HDT. It knows exactly which kind of data each sensor is producing, who are the workers online and where their data are published. In addition to that, it also knows the modules currently in use, which information they take as input and where they publish their outputs. Models are a set of descriptors defined by the administrator of the HDT that describes any worker or contextual feature.
- **Worker monitoring modules (Digital Twin Core – Data processing, analysis and decision modules):** these modules allow to elaborate data from workers, contextual sensors or any kind of system that publish data on the IoT Middleware. These modules target the detection of human status and conditions and compute complex features to allow human and machines decision-makers to consider the human factors within their execution and control logics.

The STAR's HDT has been conceived and designed to be **extensible** and **scalable**. In the STAR project, the HDT is applied in two different use cases: (i) Human-Cobot Collaboration improving robust Quality Inspections; (ii) Human Behaviour Prediction and Safety Zones Detection for Routing. For these first experiments, sensors and specific human features have been selected (e.g., Hearth Rate, accelerations, occupied space) to fulfil the expected objectives (e.g. monitor worker fatigue, predict safety zones). However, the STAR's HDT has been conceived to easily integrate new types of sensors and devices and to characterise and model several other features. This is made possible by defining a set of shared interfaces to have an easily integrable and reliable plug and play environment. In addition to the extensibility, in such kind of solution, it is also essential to consider the scalability. Different production systems involve a different number of operators, machines and sensors. The HDT must be applicable in different environments, collecting data from multiple sources, from different and numerous workers and contexts, always ensuring the same reliability.

Moreover, the STAR's HDT has been designed to be **interoperable**. It integrates Models using a modular and flexible syntax that is understandable to humans and machines alike. These Models describe how data and broadcast messages are structured, to allow all the software components to collaborate. This, together with the IoT Middleware that supports the exchange between different type of sources and destinations, allows to integrate different types of AI modules, and also to receive and share data with different types of systems, including simulation engines, PLCs and legacy ICT systems.

The HDT is indeed a cornerstone for human-aware optimization, simulation, what-if analysis, and monitoring, that are key strategic activities for manufacturing companies, in order to improve efficiency of the configuration and use of production resources. Based on massive, cumulative, real-time, real-world data, the HDT represents an evolving profile of the human-centric process in the digital world, that provides important insights on system performance and sustainability, leading to effective actions in the physical world.

7.6 Concluding Remarks

In the present literature the concept of DT is well established and numerous applications involving products, machines and equipment exist, nevertheless, only very few examples involve human aspects. To cover this gap, this work provides technological and architectural insights to realise a HDT, detailing how the presented approach is developed and used in the context of the STAR project. Future research will focus on the development of libraries to realise the HDT based on the described technological architecture, on the improvement of existent Worker Monitoring Modules and on the development of a new module, supporting decision-making within collaborative work cells.

Acknowledgements

This work has been carried out in the H2020 STAR project, which has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 956573.

References

[1] Emerton-data. AI for manufacturing (http://www.emerton-data.com/ai-for-manufacturing)

[2] World Economic Forum. (2020). The Future of Jobs Report 2020.

[3] Nunes, D. S., Zhang, P., and Silva, J. S. (2015). A survey on human-in-the-loop applications towards an internet of all. IEEE.

[4] European Commission. (2020). Industry 5.0: Towards a sustainable, human-centric and resilient European industry.

[5] Romero, D., Stahre, J., and Taisch, M. (2020). The Operator 4.0: Towards socially sustainable factories of the future.

[6] Umbrello, S., Padovano, A., and Gazzaneo, L. (2020). Designing the Smart Operator 4.0 for Human Values: A Value Sensitive Design Approach.

[7] Kaasinen, E., Schmalfuß, F., Özturk, C., Aromaa, S., Boubekeur, M., Heilala, J., … and Walter, T. (2020). Empowering and engaging industrial workers with Operator 4.0 solutions. Computers & Industrial Engineering, 139, 105678.

[8] Romero, D., Bernus, P., Noran, O., Stahre, J., and Fast-Berglund, Å. (2016, September). The operator 4.0: human cyber-physical systems & adaptive automation towards human-automation symbiosis work systems. In IFIP international conference on advances in production management systems (pp. 677–686). Springer, Cham.

[9] Zhou, J., Zhou, Y., Wang, B., and Zang, J. (2019). Human–cyber–physical systems (HCPSs) in the context of new-generation intelligent manufacturing. Engineering, 5(4), 624–636.

[10] E. Glaessgen and D. Stargel, "The digital twin paradigm for future NASA and US Air Force vehicles," in Paper for the 53rd Structures, Structural Dynamics, and Materials Conference: Special Session on the Digital Twin, 2012.

[11] Lim, K. Y. H., Zheng, P., and Chen, C. H. (2019). A state-of-the-art survey of Digital Twin: techniques, engineering product lifecycle management and business innovation perspectives. Journal of Intelligent Manufacturing, 1–25.

[12] Sengan, S., Kumar, K., Subramaniyaswamy, V., and Ravi, L. (2021). Cost-effective and efficient 3D human model creation and re-identification application for human digital twins. Multimedia Tools and Applications, 1–18.

[13] Berisha-Gawlowski, A., Caruso, C., and Harteis, C. (2021). The Concept of a Digital Twin and Its Potential for Learning Organizations. In Digital Transformation of Learning Organizations (pp. 95–114). Springer, Cham.

[14] Kawamura, R. 2019. A Digital World of Humans and Society—Digital Twin Computing. NTT Technical Review 18(3):11–17.

[15] Montini Elias, Bettoni Andrea, Ciavotta Michele, Carpanzano Emanuele, Pedrazzoli Paolo. (2021). A meta-model for modular composition of tailored human digital twins in production. 54th CIRP Conference on Manufacturing Systems.

[16] Dale, S., Kruger, K., and Basson, A. (2019, January). Human Digital Twin for Integrating human workers in Industry 4.0. In International Conference on Competitive Manufacturing, Stellenbosch University, South Africa.

[17] Z. S. Maman, M. A. A. Yazdi, L. A. Cavuoto and F. M. Megahed, "A data-driven approach to modeling physical fatigue in the workplace using wearable sensors," Applied ergonomics, vol. 65, pp. 515–529, 2017.

[18] Z. S. Maman, Y. J. Chen, A. Baghdadi, S. Lombardo, L. A. Cavuoto and F. M. Megahed, "A data analytic framework for physical fatigue management using wearable sensors," Expert Systems with Applications, 2020.

[19] V. Villani, M. Righi, L. Sabattini and C. Secchi, "Wearable Devices for the Assessment of Cognitive Effort for Human–Robot Interaction," IEEE Sensors Journal, vol. 20, no. 21, pp. 13047–13056, 2020.

[20] Peruzzini, M., Grandi, F., and Pellicciari, M. (2017). Benchmarking of tools for user experience analysis in Industry 4.0. Procedia manufacturing, 11, 806–813.

[21] Egilmez, G., Erenay, B., and Süer, G. A. (2014). Stochastic skill-based manpower allocation in a cellular manufacturing system. Journal of Manufacturing Systems, 33(4), 578–588.

[22] Graessler, I., and Pöhler, A. (2017, December). Integration of a digital twin as human representation in a scheduling procedure of a cyber-physical production system. In 2017 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM) (pp. 289–293). IEEE.

[23] Graessler, I., and Poehler, A. (2018). Intelligent control of an assembly station by integration of a digital twin for employees into the decentralized control system. Procedia Manufacturing, 24, 185–189.

[24] D'Addona, D. M., Bracco, F., Bettoni, A., Nishino, N., Carpanzano, E., and Bruzzone, A. A. (2018). Adaptive automation and human factors in manufacturing: An experimental assessment for a cognitive approach. CIRP Annals, 67(1), 455–458.

[25] Paredes-Astudillo Y A, Jimenez J F, Zambrano-Rey G, Trentesaux D. Human-Machine Cooperation for the Distributed Control of Hybrid Control Architecture. In: International Workshop on Service Orientation in Holonic and Multi-Agent Manufacturing, 98–110, 2019.

[26] Carpanzano E, Bettoni A, Julier S, Costa J C, Oliveira M. Connecting Humans to the Loop of Digitized Factories' Automation System.

In: International Conference on the Industry 4.0 model for Advanced Manufacturing, 2018.

[27] A. Bilberg and A. A. Malik, "Digital twin driven human–robot collaborative assembly," CIRP Annals, vol. 68, no. 1, pp. 499–502, 2019.

[28] Bettoni, A., Montini, E., Righi, M., Villani, V., Tsvetanov, R., Borgia, S., … and Carpanzano, E. (2020). Mutualistic and adaptive human-machine collaboration based on machine learning in an injection moulding manufacturing line. Procedia CIRP, 93, 395–400.

[29] Baskaran, S., Niaki, F. A., Tomaszewski, M., Gill, J. S., Chen, Y., Jia, Y., … and Krovi, V. (2019). Digital human and robot simulation in automotive assembly using siemens process simulate: a feasibility study. Procedia Manufacturing, 34, 986–994.

[30] Bortolini, M., Faccio, M., Gamberi, M., and Pilati, F. (2020). Motion Analysis System (MAS) for production and ergonomics assessment in the manufacturing processes. Computers & Industrial Engineering, 139, 105485.

[31] Ferraguti, F., Villa, R., Landi, C. T., Zanchettin, A. M., Rocco, P., and Secchi, C. (2020). A Unified Architecture for Physical and Ergonomic Human–Robot Collaboration. Robotica, 38(4), 669–683.

[32] X. Shen, I. Awolusi and E. Marks, "Construction equipment operator physiological data assessment and tracking," *Practice Periodical on Structural Design and Construction*, vol. 22, no. 4, 2017.

[33] T. Cheng, G. Migliaccio, J. Teizer and U. Gatti, "Data Fusion of Real-time Location Sensing (RTLS) and Physiological Status Monitoring (PSM) for Ergonomics Analysis of Construction Workers," *Journal of Computing in Civil Engineer*, 2013.

[34] L. Liu, Y. Zhang, L. Yang, L. Zhou, F. Ren and Wang, "A novel cloud-based framework for the elderly healthcare services using digital twin," IEEE Access, vol. 7, 2019.

[35] B. R. Barricelli, E. Casiraghi, J. Gliozzo, A. Petrini and S. Valtolina, "Human Digital Twin for Fitness Management," IEEE Access, vol. 8, pp. 26637–26664, 2020.

[36] Nikolaos Nikolakis, Kosmas Alexopoulos, Evangelos Xanthakis, and George Chryssolouris. The digital twin implementation for linking the virtual representation of human-based production tasks to their physical counterpart in the factory-floor. International Journal of Computer Integrated Manufacturing, 32(1):1–12, 2019.

[37] Redelinghuys, A. J. H., Kruger, K., and Basson, A. (2019, October). A six-layer architecture for digital twins with aggregation. In International Workshop on Service Orientation in Holonic and Multi-Agent Manufacturing (pp. 171–182). Springer, Cham.

[38] Steindl, G., Stagl, M., Kasper, L., Kastner, W., and Hofmann, R. (2020). Generic Digital Twin Architecture for Industrial Energy Systems. Applied Sciences, 10(24), 8903.

[39] Qian Zhu, Ruicong Wang, Qi Chen, Yan Liu, and Weijun Qin. Iot gateway: Bridging wireless sensor networks into internet of things. In 2010 IEEE/IFIP International Conference on Embedded and Ubiquitous Computing, pages 347–352. Ieee, 2010.

[40] Temesvári, Z. M., Maros, D., and Kadar, P. (2019). Review of Mobile Communication and the 5G in Manufacturing. Procedia Manufacturing, 32, 600–612.

[41] Gaston C Hillar. MQTT Essentials-A lightweight IoT protocol. Packt Publishing Ltd, 2017.

[42] Nishant Garg. Apache kafka. Packt Publishing Ltd, 2013.

[43] https://mosquitto.org

[44] Mohammad Nasar and Mohammad Abu Kausar. Suitability of influxdb database for iot applications. International Journal of Innovative Technology and Exploring Engineering, 8(10):1850–1857, 2019.

[45] Kamienski, C., Soininen, J. P., Taumberger, M., Dantas, R., Toscano, A., Salmon Cinotti, T., … and Torre Neto, A. (2019). Smart water management platform: Iot-based precision irrigation for agriculture. Sensors, 19(2), 276.

[46] Namiot, Dmitry, and Manfred Sneps-Sneppe. "On micro-services architecture." International Journal of Open Information Technologies 2.9 (2014): 24–27.

[47] Martınez-Miranda, J., and Aldea, A. (2005). Emotions in human and artificial intelligence. Computers in Human Behavior, 21(2), 323–341.

DOI: 10.1561/9781680838770.ch8

Chapter 8

Video Analytics for Situation Awareness Safe Robot-Human Cohabitation in Production Lines

By Jean-Emmanuel Haugeard and Andreina Chietera

Nowadays with the Fourth Industrial Revolution (or Industry 4.0), the automation of traditional manufacturing and industrial practices required the deployment of mobile robots that are involved to accomplish several tasks to assist workers in a modular production line. The robots are equipped with several embedded sensors (radar, camera) to analyse the nearby environment, in order to move safely and avoid obstacles. Despite, after that this technology does not provide to the robots a dynamical global view of the scene. Thus, the cohabitation between humans and robots can lead to dangerous situations. In order to ensure security between robots and workers, security zones must be detected dynamically throughout the infrastructure. For that, we will implement algorithms to analyse the scene using the global point of view of the camera network already deployed in the factory. Video analytics allows to exploit automatically the video streams in real time with the aim to detect anomalies and to raise immediately an alarm. To this end, the algorithms detect and track elements of interest (such as people, robot and new object occupying the scene) over the time, and alert the robots of the presence of any obstacles in the surrounding area. Where a human is detected close to the robot, his movements will be monitored. Based on a human behaviour analysis, the system will decide whether a new robot 'path should be calculated to reach the docking station or to stop completely to avoid any collision. This chapter presents a brief overview of these modern computer vision approaches: to detect objects of interest in video streams, and to localize them in the 3D environment. The purpose of these video analytics is to feed a "planner" indicating dynamically which areas should be avoided by a robots' fleet operating in the production lines.

8.1 Introduction

One of the main goals of STAR is to ensure the optimization of a production line to increase the efficiency of the manufacturing process. We start from the assumption that efficiency and safety go hand in hand in the complex environment of a production line, in which operators, robots and automatic systems share dynamically the same physical workspace.

The aim of this task is to take advantage of modern computer vision approaches in order to recognize postures and motion of workers and locate them as well as the items occupying the environment. The main output will be an "average spatial heatmap" representing a probabilistic occupancy of the production lines based on fixed RGB cameras deployed in the factory. The purpose of this module is to feed a "planner" indicating dynamically which areas should be avoided by the robots' fleet.

The solution we imagine is conceived by merging the following technologies:

- Dynamic object detection via Convolutional Neural Network (CNN)
- Skeleton extraction by human pose detection CNN
- 3D-localization and motion in the infrastructure and estimation of human-robot distances using the geometric calibration of fixed RGB cameras
- Heterogeneous and homogeneous multi-sensor fusion merging video analytics results coming from cameras dispatched in the production lines including other localization sensor data.

8.2 Detection of Objects of Interest in Video Streams: A Short Overview

The aim of this chapter is to highlight some video analytics approaches based on object detection and classification. In our context of monitoring for security purposes of factories, the STAR system must be able to analyse the scene and monitor the robots deployed in the factories. The aim of this video analysis module is to include in to the STAR system a software detecting empty area for secure robot displacements. The system we imagine should be able to detect the obstacles in order to avoid collision and to modify the robot planner dynamically. The objects to detect are: moving items, static object/obstacle on the navigation path and human occupying the robot's neighbourhood.

8.2.1 Moving Object Detection Using Background Modeling

The image segmentation into background regions and moving objects is a crucial stage in the video applications. The segmentation result is often used as an input for object detection/classification. Background subtraction methods are based on the premise that the difference between the background model and the current image is due to the presence of moving objects in the scene under observation.

The proposed approaches are based on background modeling of the observed scene ("background") as a first step, then on the analysis of the differences between each image and the estimated background (cf. Figure 8.1).

The foreground segmentation is possible under certain conditions:

- The camera is static (properties do not change)
- The background is statically visible most of the time
- The background is quasi-stable and can be modelled statistically over time
- Objects of interest are different (color/texture) from the background model in order to detect the difference between the current image and the background model.

A detailed survey of various background modeling methods in video analysis applications can be found in [1]. The background subtraction approaches can be divided in 4 categories:

- Basic methods : define the background as the mean or median of the observed values.
- Filtering methods (e.g. Wiener filter [2], Kalman filter [3]): design dynamic backgrounds by adaping the model using a filter.

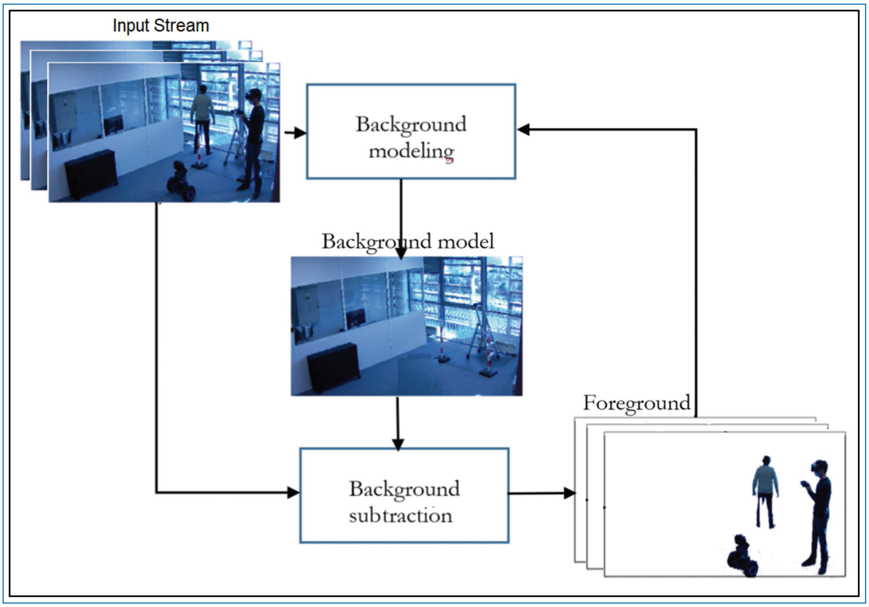


Figure 8.1. Basic steps for background subtraction algorithms.

- Clustering methods (e.g. K-Means [4], Codebooks [5]): compare the current pixel and the different clusters at every point in the image.
- Stochastic methods (e.g. Gaussian model [6], Gaussian mixture model – GMM [7], Kernel density estimation – KDE [8]): use probabilistic modeling of the background.

Stochastic methodes (GMM approaches) are more commonly used in the video applications.

The background subtraction allows to extract the "foreground" of the scene, namely the silhouettes of new or moving objects (people, vehicles, objects newly occupying the camera point of view) in the scene, but also extract areas in which lighting changes appeared due to the variations of the lighting conditions during the day. Moreover, if the objects are close to each other and/or they hide each other, their silhouettes are merged together as a single element and the resulting foreground is difficult to analyse by its shape.

This type of approach therefore makes it possible to detect all the changes in the scene, which may correspond to the presence of a new element (objects or people), but also to the presence of moving people/robot.

8.2.2 Object Detection and Classification Using CNN

Today, as in the field of image classification, object detection approaches are all based on Convolutional Neural Network architecture (CNN). These solutions based on CNN architecture consist of two parts: a “feature extractor” called backbone and a “feature classifier”. In the field of object detection based on deep learning [9], the architectures usually can be divided into two categories: two-stage and single-stage approaches.

- Two-stage detector

Two-stage networks use “Region Proposal Network” algorithm as a first step to quickly select the best candidate windows. These windows (from a few hundred to a few thousand) are then processed by a classification model (the second step) to decide whether or not they contain an object from the list considered. The most cited examples are the R-CNN model (Regions with CNN features [10]) and its derivatives: Fast R-CNN [11]), Faster R-CNN [12] and Mask R-CNN [13].

- One-stage detector

The one-stage detectors propose predicted boxes from input images directly without the region proposal step, thus they are time efficient and can be used for real-time applications. The one-stage detectors apply the classification directly to dense window grids (“anchors”) of different sizes (cf. Figure 8.2). The two main representatives of this family are the YOLO model (You Only Look Once [14]) and its derivatives: Yolov2, Yolov3 ([15]), Yolo9000, and the SSD model (Single Shot Detector [16]).

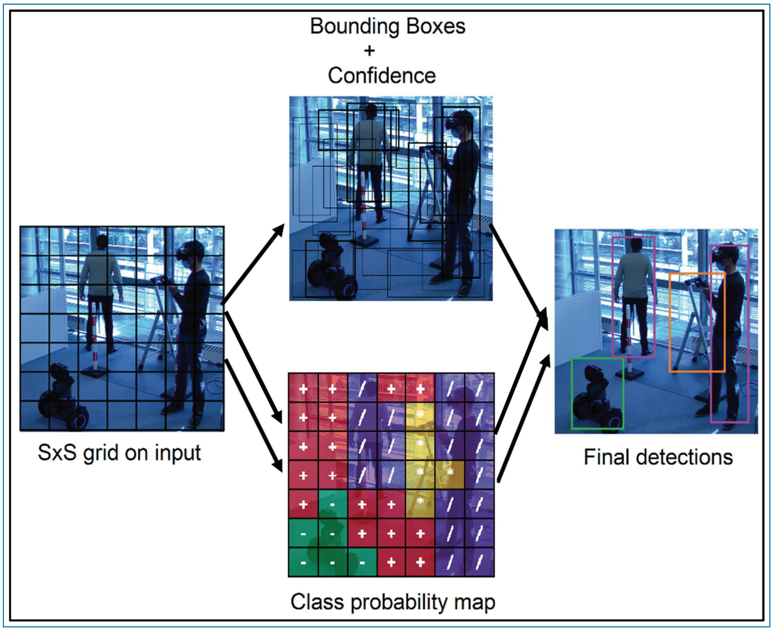


Figure 8.2. Object Classification using YOLO Algorithm in the context of Robot-Human Cohabitation. In this result, the final detection are robot, human and stool.

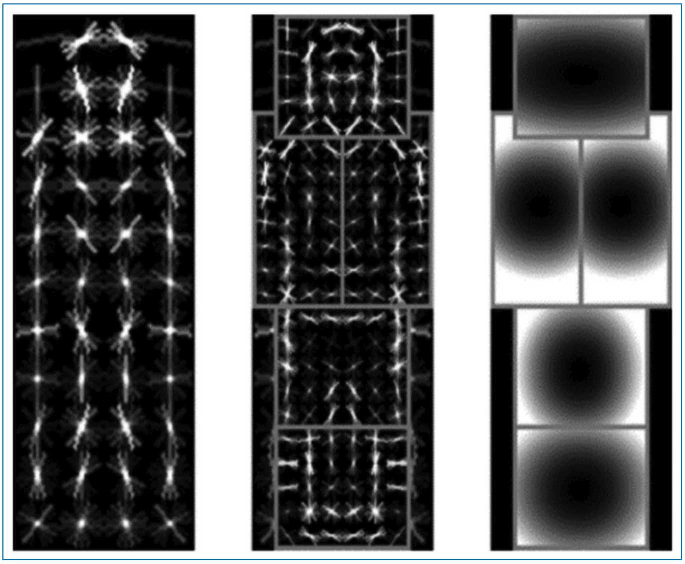


Figure 8.3. Deformable Part Model: model for the person category.

The performance of these models depends on their own architecture (meta-architecture), but also on those of the backbone used. Among the CNNs most used in this role, there are two families of state-of-the-art models for classification: VGG16 and its derivatives [17] and ResNet models (Deep Residual Learning [18]).

As a result, the vast majority of these techniques sketches, for each object detected, a rectangle called a "bounding box" surrounding the object in the image. The main exception is Mask R-CNN, which additionally provides the "mask" as the shape of each object detected, consisting in all the pixels belonging to the object in the image.

8.2.3 Human Detection Based on Deep Learning

Flexible object (e.g. a person's body) can take multiple appearances in the image. This characteristic makes the task of detection/classification more complex. From the 2010 s, research laboratories worked on methods based on the shape of objects of interest merged with machine learning techniques to be able to take into account all the possible configuration of the shape (feature templates – Deformable Part Model (DPM) [19] Figure 8.3). These techniques relied on the use of local attributes (descriptors) such as Histogram of Oriented Gradients (HOG [20] Figure 8.4), and could be a stand-alone solution or could be applied in combination with a background subtraction method to decrease false negatives. This learning-based approach has seen significant improvements with the advent of Convolutional Neural Network CNNs, and their adaptation to object detection. The main

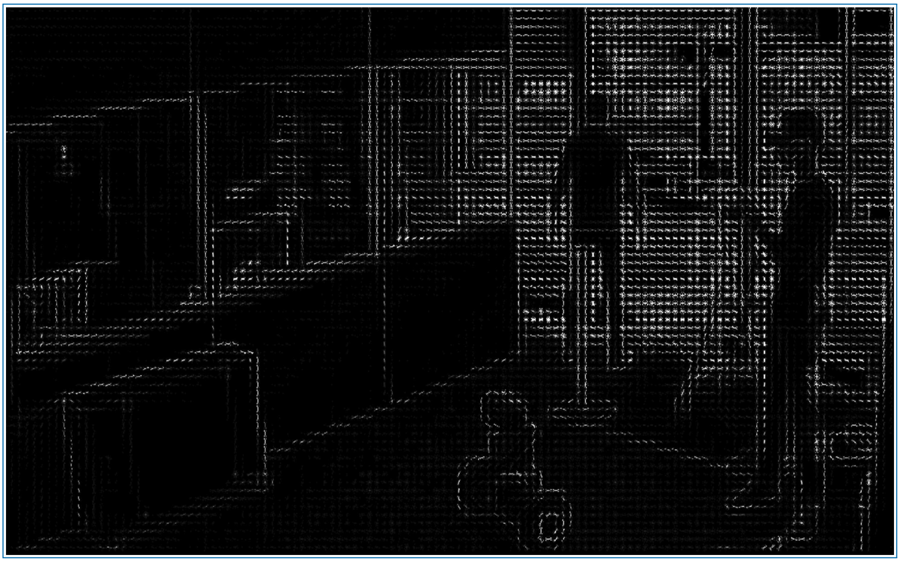


Figure 8.4. Example of HOG feature on STAR project image.

problem with the techniques proposed in the factory context is the lack of robustness when partial occlusion occurred. Indeed and especially in a production line, the occlusion affects the people detection making the task more complex.

A technique for people detection, called OpenPose, was recently proposed [21] which takes into account both the variability of the shapes observed (due to the fact that people are articulated objects) and the presence of partial occlusions. OpenPose is based on a CNN architecture and makes it possible to detect different characteristic points of the human body (joints, eyes, mouth, nose, ears, hands, feet) and, jointly, to group these points in a graph forming a skeleton representation (cf. Figure 8.5). More specifically, the skeleton detection algorithm allows to track human poses by detecting and estimating the position of the characteristic points defining human postures. The approach creates heat maps for joint extraction and extracts affinity fields considering all the detected joints in order to infer the link between them and, consequently, allow the detection of human limbs. The algorithm can simultaneously process different observation scales. It should also be noted that it can detect people both by their silhouette when it is clearly visible and by their head, which is more rarely masked.

8.2.4 Object Geolocation Using 3D Calibration

Once humans or other items are detected from video footprint, they are located in the infrastructure. This absolute positioning of the elements of the scene requires a camera calibration phase, in order to associate to each pixel of the image space the

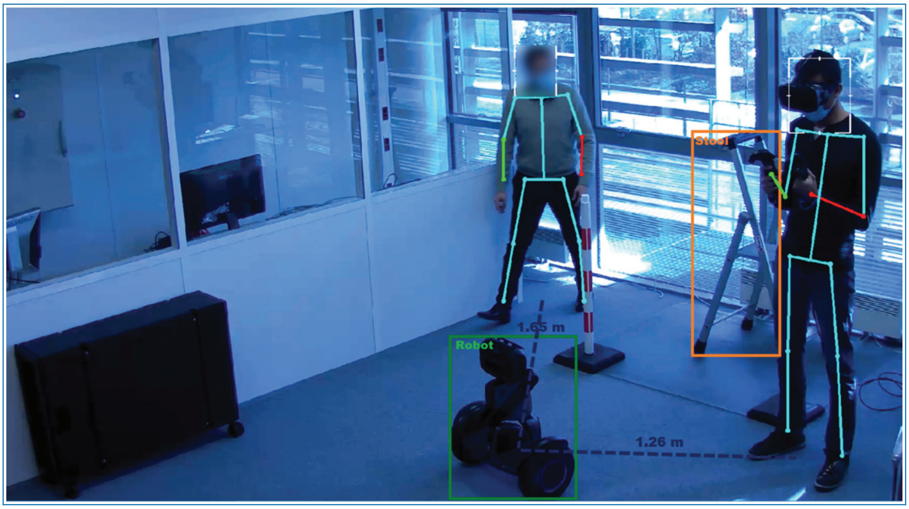


Figure 8.5. An example of the results obtained using different detectors: (1) moving object detection (GMM subtraction), (2) object detector to identify the stool and the robot (Yolo), (3) human detector (OpenPose).

coordinate in an absolute 3D coordinates system. Once an element is detected, it is projected on the ground, taking into account a reference measurement (such as the height of the body, the robot dimension). The projection on the ground allows to estimate the actual 3D position and then the distances between any other elements of the images (cf. Figure 8.5).

8.3 Conclusion

We have introduced a short overview of detection of elements in the scene with the purpose to define empty spaces for robot navigation. Based on these machine learning solutions, we will develop new innovative approaches able to analyse the global scene, alert the workers to potential danger and feed the robot path planner. In the context of STAR, the robot should be able to detect the obstacles in order to avoid collision. These video analytics will provide the input in real-time to feed a “planner” with the areas that should be avoided.

Acknowledgements

This work has been carried out in the H2020 STAR project, which has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No. 956573.

References

[1] Y. Xu, J. Dong, B. Zhang and D. Xu, "Background modeling methods in video analysis: A review and comparative evaluation," *CAAI Transactions on Intelligence Technology*, pp. 43–60, 2016.

[2] K. Toyama, J. Krumm, B. Brumitt and B. Meyers, "Wallflower: Principles and practice of background maintenance," *Proceedings of the seventh IEEE international conference on computer vision*, vol. 1, pp. 255–261, 1999.

[3] C. Ridder, O. Munkelt and H. Kirchner, "Adaptive background estimation and foreground detection using kalman-filtering," *Proceedings of International Conference on recent Advances in Mechatronics*, pp. 193–199, 1995.

[4] S. Indupalli, M. A. Ali and B. Boufama, "A novel clustering-based method for adaptive background segmentation," *The 3rd Canadian Conference on Computer and Robot Vision*, 2006.

[5] K. Kim, T. H. Chalidabhongse, D. Harwood and L. Davis, "Real-time foreground–background segmentation using codebook model," *Real-time imaging*, vol. 11, no. 3, pp. 172–195, 2005.

[6] N. Friedman and S. Russell, "Image segmentation in video sequences: A probabilistic approach," *arXiv preprint arXiv:1302.1539*, 1997.

[7] C. Stauffer and W. E. L. Grimson, "Adaptive background mixture models for real-time tracking," vol. 2, pp. 246–252, 1999.

[8] A. Elgammal, D. Harwood and L. Davis, "Non-parametric model for background subtraction," *European conference on computer vision*, pp. 751–767, 2000.

[9] J. Licheng, Z. Fan, L. Fang, Y. Shuyuan, L. Lingling, F. Zhixi and Q. Rong, "A Survey of Deep Learning-Based Object Detection," *Institute of Electrical and Electronics Engineers (IEEE)*, vol. 7, 2019.

[10] R. Girshick, J. Donahue, T. Darrell and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 580–587, 2014.

[11] R. Girshick, "Fast R-CNN," *IEEE International Conference on Computer Vision*, pp. 1440–1448, 2015.

[12] S. Ren, K. He, R. Girshick and J. Sun, "Faster R-CNN: Towards realtime object detection with region proposal networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, vol. 39, pp. 1137–1149, 6.

[13] K. He, G. Gkioxari, P. Dollar and R. Girshick, "Mask R-CNN," *IEEE International Conference on Computer Vision (ICCV)*, 2017.

[14] J. Redmon, S. Divvala, G. R. and A. Farhadi, "You only look once: Unified, real-time object detection," *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 779–788, 2016.

[15] J. Redmon and A. Farhadi, "Yolov3: An incremental improvement," *arXiv:1804.02767*, 2018.

[16] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu and A. C. Berg, "SSD: Single shot multibox detector," *European Conference on Computer Vision (ECCV)*, pp. 21–37, 2016.

[17] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *International Conference on Learning Representations (ICLR)*, 2015.

[18] K. He, X. Zhang, S. Ren and J. Sun, "Deep residual learning for image recognition," *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016.

[19] P. F. Felzenszwalb, R. B. Girshick, D. McAllester and D. Ramanan, "Object detection with discriminatively trained part-based models," *IEEE transactions on pattern analysis and machine intelligence*, vol. 32, no. 9, pp. 1627–1645, 2009.

[20] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," *IEEE Conference on Computer Vision and Pattern Recognition*, 2005.

[21] Z. Cao, G. Hidalgo, T. Simon, S.-E. Wei and Y. Sheikh, "OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields," *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

DOI: 10.1561/9781680838770.ch9

Chapter 9

Human in the Loop of AI Systems in Manufacturing

By Christos Emmanouilidis and Sabine Waschull

Artificial Intelligence (AI) in manufacturing is typically looked upon from the viewpoint of its contribution to automation. Additionally, the role of AI in augmenting human activities has been the subject of a wide range of studies with impact on practical applications in manufacturing environments. Recently, the empowering effect of human and AI actors working in synergy has attracted increased attention. After outlining relevant work, this chapter considers the potential emergent outcomes of such a synergy in a way that goes beyond automation or augmentation. Aimed at both developers and work designers, the present work proposes a model of human-AI interaction along with an outline of key concepts and success criteria towards making human-AI interaction more effective.

9.1 Introduction

Early work on the integration of human and technical actors in sociotechnical systems has given rise to the field of Human System Engineering (HSE), defined as the application of principles, models, and techniques to system design, taking into account human capabilities and limitations [1]. The HSE field has evolved substantially, leading to significant advances in Human-Systems Integration (HSI). While HSI has matured to take into account automation and ergonomics considerations, it has been far less concerned with the integration of Artificial Intelligence (AI). Yet, the integration challenges associated with introducing AI elevates the need to consider the joint optimisation of the human and technical systems capabilities to a level that incorporates the potential outcomes of humans and AI agents acting in synergy. A sharp contrast between human and non-human actors is that of the understanding and intentionality of actions: people should understand the purpose of actions, whereas technical system target at best to perform "as instructed". This "as instructed" can be broad enough to encompass different instruction subjects: designers, operators, or even programs. Automation has empowered technical actors to create, process, and execute complex "instructions" in highly efficient ways. AI raises significantly more the capabilities of automation systems to handle or respond to siutations that automation alone would not suffice to handle but often still lacks the versatilty of human congitive capabilities to deal with uncertain, incomplete, or generally less well-defined contexts. As a result, human involvement on such tasks is valuable. However, human actions are typically more effective when human operators are acting within a shared understanding of the work activities context and this "situational awareness" is recognised in a "collective activity" view of work environments [2]. According to such a view, "collective activity" should not be viewed upon as the sum of individual activities, but through the evolving interaction of the actors which contribute to it. Therefore, when considering the interaction between human and non-human actors (including AI) in manufacturing environments, it is not sufficient to consider how human actions can be augmented by technical actors or how technical systems can be aided by humans. Instead, it is more relevant to analyse and understand the emergent outcomes of their interaction. While the interaction between human actors and automation agents has been the subject of numerous research studies, previous works mostly consider how AI supports humans [3, 4]. Beyond this, AI has recently been considered through the situational awareness perspective, as a means of collective activity effectiveness, mostly through seeking explainability and transparency in its outcomes in the form of eXplainable AI (xAI) [5]. As the introduction of Industry 4.0 technologies has reshaped work roles in profound ways [6], the collective activity viewpoint is appropriate for the understanding, analysis and design of the human-AI interaction in manufacturing. The need to consider human-centricity at the design stage of sociotechnical systems should be extended to the design of human-centred AI when considering manufacturing workplaces [7]. Starting from the empowering effect that the interaction between technical and human actors can have [8], a collective activity perspective can place this empowerment in the context of action affordances. The term "affordance" denotes "action possibilities provided to the actor by the environment" [9]. But such possibilities have a relational nature, i.e. they refer to interaction possibilities relevant to a specific actor operating in a specific environment. This renders the augmentation viewpoint too simplistic. Incorporating human activities as steps in a broader AI-driven process, termed as "human in the AI loop", implies more than augmenting the algorithm [10]. Although the physical and cognitive support capabilities offered by Industry 4.0 technologies to human workers are acknowledged [11], the actual integration of human cognitive capabilities in the AI loop is less well understood [12]. Furthermore, in most cases the AI outcomes are characterised by a lack of transparency and explanation and so, being poorly understood, they are not sufficiently trusted by humans. Therefore a shared context through situational awareness in human-AI interaction is hardly established, limiting the effectiveness of human-AI integration.

The aim of this chapter is to make a contribution towards more effective human-AI integration in manufacturing environments by looking at the potential contribution of the Human in the AI loop and then seeking to translate this to recommendations for the effective integration of humans and AI from a work design perspective. This takes the form of a conceptual model of human-AI interaction, which was put to test in co-creation workshops where different stakeholders contributed to the design of human-centric solutions for manufacturing lines. The co-creating stakeholders worked through different scenarios to produce a synthesis of interaction possibilities into new designs, while stating also expectations for the desired outcomes of the new approaches. The proposed model and co-creation design practice can be of practical value to system and work designers alike, when seeking to integrate humans and AI in human-centric deployments in manufacturing environments. The rest of the chapter is structured as follows. Section 9.2 outlines how human and AI actors interact in different AI processes and highlight the potential benefits. Section 9.3 introduces a conceptual model of interacting human and non-human (AI) actors in production environments. Section 9.4 considers work design implications of such collective activity. Section 9.5 presents the outcomes of placing both the conceptual model and the work design implications to test in a stakeholders co-creation workshop setting. Section 9.6 is the conclusion.

9.2 Humans and AI in Sociotechnical Production Systems

There is barely a single definition of what constitutes AI, but to the extent that intelligence characteristics are associated with thought processes and behaviours, the expectations for an AI agent would be to exhibit at least some of those characteristics. The thought processes are typically looked upon from the cognitive systems and logic viewpoints, while the behavioural ones may result from applying concepts, methods, and practice related to machine learning, knowledge representation and reasoning, natural language processing, and agent-based systems [13]. Despite the potential of AI to take on human tasks, e.g. the automation of physical, cognitive, discretionary and decision-making tasks [4], there is a growing consensus among researchers and practitioners to design human-centric technologies which integrate rather than eliminate humans and their capabilities [12, 14, 15].

While the majority of such human-in-the-loop scenarios consider how AI augments humans [3], the opposite (i.e. humans aiding AI) also holds significant potential for the successful joint integration of humans and AI agents in production environments [10, 12]. The advances made in the practical application of AI, involving scenarios of automation and augmentation of human work [4], create the need to better understand the interactions between human and technical actors in AI-based production environments. Human augmentation has received extensive attention in manufacturing, both from the viewpoint of technology enablers for such augmentation, as well as regarding the functional and domain-specific outcomes of the augmentation. Enabling technologies for human augmentation in manufacturing include web-services for ubiquitous computing [16], multimodal interfaces [17], augmented [18] and virtual [19] reality, context-adaptive computing [20], exoskeletons for physical augmentation [21], and natural interfaces, including speech [22] and brain – computer interfaces [23].

There are multiple types of activities wherein human actors can aid AI, however reported implementations in manufacturing environments have been far fewer compared to human augmentation [12]. Yet, the potential contribution of humans towards AI agents [10] can be very influential and valuable, even when applied to the most data-driven part of an AI process, that of machine learning [12].

9.3 Meta-Human Learning in Sociotechnical Systems

Consistent with the concept of collective activities, this section introduces a model of collective or meta-human learning, where the interest is not confined to what a specific human actor or a specific AI-driven one can learn. Instead the interest is in the emergent learning capabilities of the broader sociotechnical system.

A simplistic view is of human and AI actors' interaction is that human and non-human actor capabilities are static and so do their affordances in given technical environments. The superiority of human cognitive capabilities over AI in performing cross-domain activities is not a controversial statement and likewise the superiority of AI in data-intensive tasks has been demonstrated, leading to the definition of a range of cognitive systems architectures [24]. Efforts to bridge the deficiency of AI to perform only within narrow contexts have been mostly in transfer learning [25], aiming to transfer the learned capabilities from the original domain of the learning to a new one. There have been various examples of integrating human knowledge to machine learning [12, 26] but this is only one aspect of the integration. There is also a motivation to examine how to enable human and non-human actors benefitting from each other's capabilities and the empowering effect that they can have on each other [8]. To this end, a conceptual model for the collective contribution of such actors, acts as a meta-human learning system [27]. The term meta-human learning refers to the emergent "learning" capabilities of the overall sociotechnical system. For example, an AI-driven system that is embedded in process monitoring can be effective in a range of tasks, but may not have been trained to perform well on others. Rather than attempting to offer loosely founded and potentially erroneous outcomes, the AI-driven monitoring system may flag out cases not seen before. An expert user may be prompted to assign these cases to existing concepts or states, thus expanding the knoweldge domain of the original AI model [28]. Such an interaction process has become popular and is recognised as a form of Active Learning [29–31]. The connection of cases to concepts may actually be the product of collective user interaction and annotation of cases, a process which can be seen as a joint Linked Data and Knowledge Management, which can be instantiated on knowledge graph constructs [32]. Starting from such concepts about humans and AI interaction proposed in [8], and incorporating ideas about introducing the human cognitive capabilities in the AI loop [12], the model proposed in this work is illustrated in Figure 9.1.

Human actors, capabilities and interaction affordances within the sociotechnical manufacturing environment are marked in green. Technical actor capabilities, including both existing operational and information technology, as well as those of AI actors are marked with blue. All actors exhibit certain capabilities which can be expressed in a range of interaction affordances given the context of the operating sociotechnical environment. However, technical actors empower humans to expand their capabilities, inform them about relevant process status or knowledge, train them on certain tasks, explain (through AI) automated outcomes or recommendations, but also constraint their affordances within a range of admissible

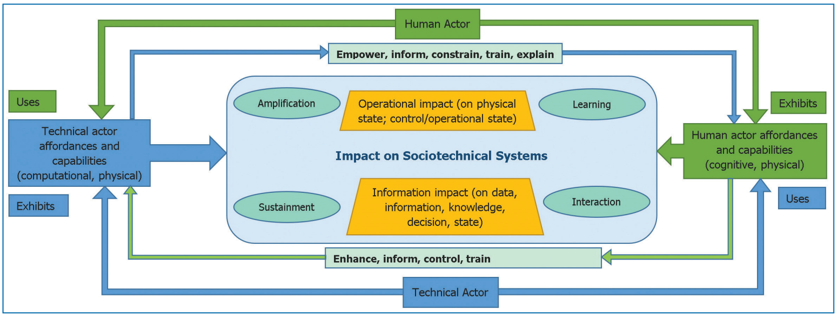

Human Actor
Empower, inform, constrain, train, explain
Uses
Technical actor affordances and capabilities (computational, physical)
Exhibits
Amplification
Operational impact (on physical state; control/operational state)
Learning
Impact on Sociotechnical Systems
Sustainment
Information impact (on data, information, knowledge, decision, state)
Interaction
Exhibits
Human actor affordances and capabilities (cognitive, physical)
Uses
Enhance, inform, control, train
Technical Actor

Figure 9.1. Meta-human learning through collective activity of actors.

actions. These are concepts that will be further explored in the industrial use cases of the STAR project.

9.4 Meta-Human Learning Considerations for Work Design

Earlier sections argued in support of Human and AI actors engaging in a collective activity, but such engagement may also give rise to different cognitive and mental demands on humans, as well as a change in the way physical activities are performed. How exactly these are likely to affect the overall performance of the operations system remains an open question [33]. It is important to design the interaction in such a way that the resulting work characteristics lead to positive outcomes for the workers and the organization at large.

The earlier discussion raises questions about how to effectively design such an integration to deliver improved performance [34]. To answer this question, it is necessary to take a work design viewpoint. The physical, cognitive and mental demands for workers may affect the overall performance of the operations system [33]. Various streams of work design theory came together in the seminal work of [35]. An overview is given in [36], including integrative perspectives that provide links between the earlier streams. Work design theory provides a set of work characteristics that should be considered when (re)designing jobs in response to technological and social changes to achieve different individual and organizational purposes. As such, it is important that the adoption of AI in industry pays due attention to these characteristics, so as to deliver a design approach for AI and humans integration and produce positive outcomes for the individuals and the organization at large. The focus is on the work characteristics that arise from specific task environment characteristics and social environment ones, and exclude those related to the broader physical and organizational environment (contextual characteristics). The terminology is taken from [37].

An influential task characteristic is **autonomy**. Autonomy refers to the amount of freedom that a human has during the work in terms of timing of the work, choice of methods, and the ability to make decisions. Jobs that lack autonomy are considered poorly designed. AI may impact autonomy in positive and negative ways [7]. **Task variety** considers the range of tasks that humans need to perform in their job, while **skill variety** relates to the required skills to perform the job. AI may replace routine cognitive tasks, but also create new tasks, requiring new skills from humans who are interacting with the system. The task and skills variety should match the abilities and needs of the individual worker. The same holds for **job complexity**: too little and the job is without challenges; too much creates fatigue and stress. AI may impact job complexity by altering the cognitive demands. **Feedback from the job**, i.e. being able to evaluate the quality of work while it is being performed, is another task characteristic. In general, AI may improve the feedback from the job due to sensor technology and visual devices that can provide such feedback. AI may contribute substantially by providing more intelligence to the feedback. Conversely, a poor division of task between the AI agent and the human agent may also lead to decreased opportunities for learning and impaired situational awareness. **Specialization** refers to the extent to which a job involves the performance of tasks requiring specific knowledge and skill, and AI may empower humans to take on a variety of tasks by supplementing knowledge and enhancing capabilities, but it may also shift human work to focus on a narrow set of specialized tasks. **Problem solving** in the job is again a task characteristic which should be challenging, but not too challenging for the individual employee. AI can execute routine problems, and create new complex problems for humans to focus on. Finally, **information processing** is a task characteristic which should match the cognitive capabilities of the worker, and which is highly influenced by digitization in general.

There are also some relevant work characteristics related to the social environment that may be impacted by the adoption of AI. Traditionally, these characteristics reflect the relations among workers. However, in modern manufacturing environments where Industry 4.0 technologies such as AI are applied in production systems, these concepts may also relate to interactions between humans and AI agents. **Interdependence** traditionally reflects to the extent that humans connect to each other, but may be expanded to also reflect the connection between humans and AI actors. Integrating the human in the AI-loop implies an increased dependency between both actors. Similarly, AI may facilitate **social support** by providing valuable connections between team members and enhancing their communication. Similar effects may be expected for the enhancement of the amount of feedback from other humans (peers or supervisors). To summarize, digitization, and specifically AI, can have both positive and negative impacts on many work characteristics. Therefore, its impact should be carefully considered in a human-centric AI design.

9.5 Co-Creation Workshops

The achievement of success criteria such as human-centricity, safety and trust has been defined as key success criteria for the development of AI technologies in the STAR project. To drive their development, and measure the successful achievement of aspired success criteria, a series of co-creation workshops were organised. In participatory design processes users play an active role in all phases of the project by proposing ideas and providing suggestions. As opposed to solely assessing the software artifact, users actively and directly participate in the design process activities through shared experimentation, mutual learning and reflection [38, 39]. Within STAR, each pilot demonstration site will conduct a number of co-design sessions throughout the design, development and testing process of the AI technology, facilitating participation to a feasible extent. By means of multiple methods and tools, input and feedback will be collected and synthesized for relevant stakeholders. The project co-ordinator's agile development process will ensure that resulting requests for changes of the software are incorporated at the interim and final release. To increase the studies validity, each session will include multiple stakeholders to assess and validate the technical artifacts throughout their development cycle. Workshops have been planned for both the definition and design, as well as the two development and testing phases at each pilot site. The development and testing phase workshops will contain an evaluation part (focus groups) and a co-creation part. An overview of the workshops and their goals can be seen in Figure 9.2.

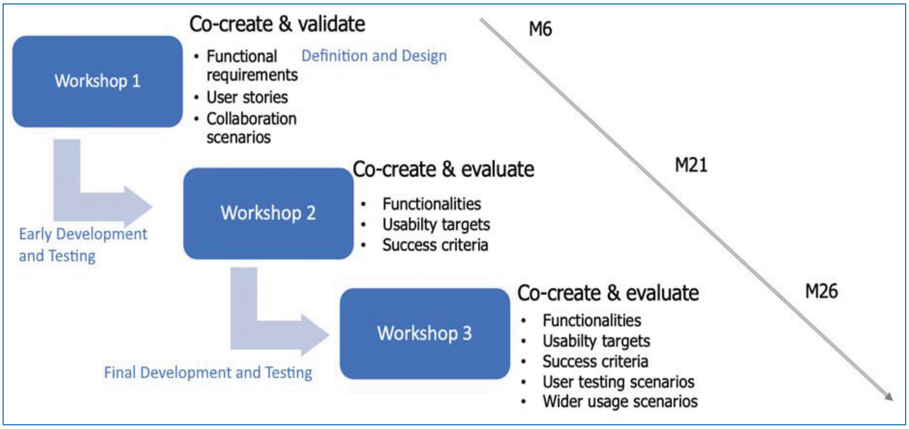


Figure 9.2. STAR co-creation workshops.

Definition and Design Phase Workshops (W1)

W1 workshops have been planned for the definition and design phase of the AI technology and the pilot user scenarios. In the first part of the workshop, participants are asked to validate and evaluate the functional and non-functional requirements and user stories in a focus group, and then in the second part they co-create different collaboration scenarios based on a pre-defined (but open to additions or modifications) relevant success criteria. Scenarios address approaches regarding how (1) humans can help/augment the AI, (2) where AI technology can help/augment humans, (3) where AI substitutes humans, and (4) where AI and humans are integral part of a process, with no clear unidirectional support for each other.

Early Development and Testing Phase Workshops (W2)

During W2, co-creation and evaluation activities will be undertaken addressing the functionalities of the first version of the pilot systems available (Early Design and Development). Participants will visualize, simulate and experiment with the pilot system, supported by mock-ups and prototypes. Based on that, participants can develop and propose improvement ideas and system testing scenarios. Moreover, the usability targets and success criteria will be evaluated. The workshop is planned after the first iteration of STAR components and systems have been deployed at pilot sites (Interim version prototype implementation of pilot systems).

Final Development and Testing Phase Workshops (W3)

The content and focus of Workshops W3 are similar to W2, but the co-creation and evaluation activities will be focused on the final version of the pilot systems. In addition to the aims of W2, the final workshops will additionally propose final user testing scenarios, as well as usage scenarios for wider stakeholders external to the project. Workshops W3 will be planned to provide sufficient impetus and time for acting upon their outcomes, a few months before the delivery final prototype implementation of the pilot systems.

An outline of the co-creation workshop targets can be seen in Table 9.1. Specifically for the W1 workshops, the initial pool of success criteria were defined by taking into account both pilot targets and work design characteristics, as well as a survey among STAR project partners, as a representative expression of a wider stakeholders view involving industrial end users, technology providers, legal and ethics experts, and research organisations. These success criteria can be seen in Figure 9.3. Each workshop was "seeded" with initial user stories, potential components and desired functionalities to satisfy the user stories, as well as initial list of tasks relevant to the role of humans and AI, and were accompanied by a mapping of the process workflow under consideration in each pilot use case. Co-creation workshops have

Table 9.1. Co-creation workshop targets.

Co-creation Workshops	Workshop Target
Definition and Design Workshops (W1)	• Validate requirements • Validate user stories • Co-create task scenarios
Early Development and Testing Workshops (W2)	• Co-create testing scenarios • Obtain early feedback from stakeholders (testing and validation)
Final Development and Testing Workshops (W3)	• Final testing and validation scenarios • Assess future prospects and suggest areas for further development within and beyond the project

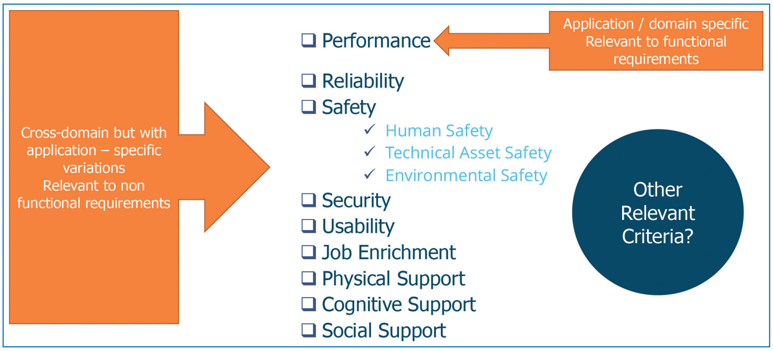


Figure 9.3. Pilots co-creation workshops initial pool of success criteria.

already been conducted for two of the three pilot cases regarding the definition and design stage. However, as this is an ongoing activity, the present work only includes preliminary information, as the full synthesis of the outcomes has been planned for the next period. Due to the COVID-19 pandemic restrictions, the co-creation workshops were held using the MIRO[1] collaboration tool. For each workshop, initial collaboration boards were set up, breaking down the collaboration activities to the following stages.

1. User stories definition and adaptation
2. Functional requirements and components definition
3. Linking user stories with functional requirements and components
4. Classifying Human – AI activities into sub-categories: (i) AI augments humans (ii) humans augmenting AI (iii) AI substitutes humans (iv) human-AI integrated tasks

[1] https://miro.com/.

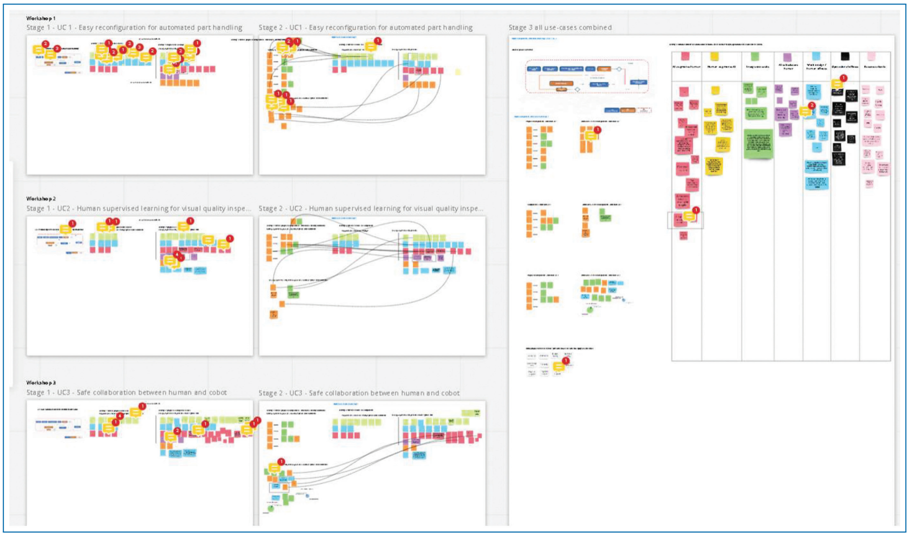

Workshop 1
Stage 1 - UC 1 - Easy reconfiguration for automated part handling
Stage 2 - UC1 - Easy reconfiguration for automated part handling
Stage 3 all use-cases combined
Workshop 2
Stage 1 - UC2 - Human supervised learning for visual quality inspe...
Stage 2 - UC2 - Human supervised learning for visual quality inspe...
Workshop 3
Stage 1 - UC3 - Safe collaboration between human and cobot
Stage 2 - UC3 - Safe collaboration between human and cobot

Figure 9.4. Co-creation collaboration board activity.

5. Drafting work design/human effects, as well as operational effects and success criteria (following task characteristics mentioned in section 4 and as seen in Figure 9.3)

An example of a MIRO collaboration activity can be seen in Figure 9.4. During the co-creation workshops the collaborating partners were able to produce an enhanced and expanded canvas of user stories, components, and design elements regarding human-AI synergies and expected outcomes. However, the full analysis of the co-creation workshop outcomes will be part of a forthcoming report and publication.

9.6 Conclusion

This chapter considered the interaction between human and non-human actors in industrial environments. It analysed the nature and benefits arising from this interaction and highlighted that AI-enabled manufacturing environments do not just benefit from better performing humans or machines, but also from their expanded capabilities. To unleash the interaction benefits, design approaches for the effective integration of human and AI actors in manufacturing are needed. This necessitates interdisciplinary synergies involving manufacturing operations, AI, as well as work design for Industry 4.0. To this end, the paper presented a synthesis of key synergies between human and AI-actors, it proposed a model of emergent meta-human learning in sociotechnical systems and provided directions for a work design viewpoint for such an effective integration. As part of the STAR collaborative research project, which integrates manufacturing industries, technology providers, research organisations, and legal/ethics stakeholders, current research analyses industrial case studies involving industrial assembly and quality inspection, agile production, and human cobot collaboration in Industry 4.0 environments. Furthermore, it examines industrial requirements and success criteria, including overall operational performance, technical system components performance, and human and job effects of AI. The studied human and job effects are an elaboration of factors considered in Section 9.4, while aspects of human – AI interactions, discussed in Sections 9.2 and 9.3, as also assessed. The reported work has placed the initial findings as "seeds" for a series of co-creation workshops which have been designed to take place at both the design, as well as the implementation and testing faces of the STAR system and its components. Among the co-creation seeds were an initial pool of success criteria for the overall sociotechnical system, such as reliability, performance, safety (human, technical, environmental), security, usability, worker support (physical, cognitive, social), and job enrichment, to determine an effective design approach for the integration of humans and AI in production environments. Further work will progress with the full analysis and synthesis from the design phase co-creation workshops. The outcomes of this analysis are seen as valuable input for the STAR project design approach for trusted and safe human-centric AI in manufacturing environments.

Acknowledgements

This work has been carried out in the H2020 STAR project, which has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 956573.

References

[1] DOD, "Manpower, personnel, training, and safety (MPTS) in the defense system acquisition process. DoD Directive 5000.53," Washington, DC, 1988.

[2] S. Caroly and F. Barcellini, *A conceptual framework of collective activity in constructive ergonomics*, vol. 822. Springer International Publishing, 2019.

[3] R. Raisamo, I. Rakkolainen, P. Majaranta, K. Salminen, J. Rantala, and A. Farooq, "Human augmentation: Past, present and future," *Int. J. Hum. Comput. Stud.*, vol. 131, no. May, pp. 131–143, 2019, doi: 10.1016/j.ijhcs.2019.05.008.

[4] S. Raisch and S. Krakowski, "Artificial intelligence and management: The automation–augmentation paradox," *Acad. Manag. Rev.*, vol. 46, no. 1, pp. 192–210, 2021, doi: 10.5465/AMR.2018.0072.

[5] I. Seeber, L. Waizenegger, S. Seidel, S. Morana, I. Benbasat, and P. B. Lowry, "Collaborating with technology-based autonomous agents: Issues and research opportunities," *Internet Res.*, vol. 30, no. 1, pp. 1–18, 2020, doi: 10.1108/INTR-12-2019-0503.

[6] W. P. Neumann, S. Winkelhaus, E. H. Grosse, and C. H. Glock, "Industry 4.0 and the human factor – A systems framework and analysis methodology for successful development," *Int. J. Prod. Econ.*, vol. 233, no. September 2020, p. 107992, 2021, doi: 10.1016/j.ijpe.2020.107992.

[7] S. K. Parker and G. Grote, "Automation, Algorithms, and Beyond: Why Work Design Matters More Than Ever in a Digital World," *Appl. Psychol.*, vol. 0, no. 0, pp. 1–45, 2020, doi: 10.1111/apps.12241.

[8] H. James Wilson and P. R. Daugherty, "Collaborative intelligence: Humans and AI are joining forces," *Harv. Bus. Rev.*, vol. 2018, no. July–August, pp. 114–123, 2018.

[9] V. Kaptelinin, B. Nardi, B. Hall, and U. C. Irvine, "Affordances in HCI: Toward a Mediated Action Perspective," pp. 967–976, 2012.

[10] T. Grønsund and M. Aanestad, "Augmenting the algorithm: Emerging human-in-the-loop work configurations," *J. Strateg. Inf. Syst.*, vol. 29, no. 2, p. 101614, 2020, doi: 10.1016/j.jsis.2020.101614.

[11] C. Cimini, F. Pirola, R. Pinto, and S. Cavalieri, "A human-in-the-loop manufacturing control architecture for the next generation of production systems," *J. Manuf. Syst.*, vol. 54, no. July 2019, pp. 258–271, 2020, doi: 10.1016/j.jmsy.2020.01.002.

[12] C. Emmanouilidis *et al.*, "Enabling the human in the loop: Linked data and knowledge in industrial cyber-physical systems," *Annu. Rev. Control*, vol. 47, pp. 249–265, 2019, doi: 10.1016/j.arcontrol.2019.03.004.

[13] S. Russel and P. Norvig, *Artificial Intelligence: A Modern Approach*. Pearson, 2020.

[14] B. A. Kadir and O. Broberg, "Human-centered design of work systems in the transition to industry 4.0," *Appl. Ergon.*, vol. 92, no. November 2020, p. 103334, 2021, doi: 10.1016/j.apergo.2020.103334.

[15] D. Romero *et al.*, "Towards an operator 4.0 typology: A human-centric perspective on the fourth industrial revolution technologies," *CIE 2016 46th Int. Conf. Comput. Ind. Eng.*, no. April 2017, pp. 0–11, 2016.

[16] M. Lampe, M. Strassner, and E. Fleisch, "A Ubiquitous Computing environment for aircraft maintenance," *Proc. 2004 ACM Symp. Appl. Comput. – SAC '04*, p. 1586, 2004, doi: 10.1145/967900.968217.

[17] C. Washburn, P. Stringfellow, and A. Gramopadhye, “Using Multimodal Technologies to Enhance Aviation Maintenance Inspection Training,” *Digital Human Modeling*, no. 4561, pp. 1018–1026, 2007, doi: https://doi.org/10.1007/978-3-540-73321-8_114.

[18] B. Schwald and B. DeLaval, “An Augmented Reality System for Training and Assistance to Maintenance in the Industrial Context,” *11th Int. Conf. Cent. Eur. Comput. Graph. Vis. Comput. Vis.*, pp. 425–432, 2003, doi: 10.1007/11941354_29.

[19] J. R. Li, L. P. Khoo, and S. B. Tor, “Desktop virtual reality for maintenance training: An object oriented prototype system (V-REALISM),” *Comput. Ind.*, vol. 52, no. 2, pp. 109–125, 2003, doi: 10.1016/S0166-3615(03)00103-9.

[20] N. Papathanasiou, D. Karampatzakis, D. Koulouriotis, and C. Emmanouilidis, “Mobile Personalised Support in Industrial Environments: Coupling Learning with Context – Aware Features,” *IFIP Adv. Inf. Commun. Technol.*, vol. 438, no. PART 1, pp. 298–306, 2014, doi: 10.1007/978-3-662-44739-0_37.

[21] S. Fox, O. Aranko, J. Heilala, and P. Vahala, “Exoskeletons: Comprehensive, comparative and critical analyses of their potential to improve manufacturing performance,” *J. Manuf. Technol. Manag.*, vol. 31, no. 6, pp. 1261–1280, 2020, doi: 10.1108/JMTM-01-2019-0023.

[22] S. Goose, S. Sudarsky, X. Zhang, and N. Navab, “Speech-enabled augmented reality supporting mobile industrial maintenance,” *IEEE Pervasive Comput.*, vol. 2, no. 1, pp. 65–70, 2003, doi: 10.1109/MPRV.2003.1186727.

[23] B. Zhang, J. Wang, and T. Fuhlbrigge, “A review of commercial brau=in computer intergaces from the perspective of industrial robotics,” in *Proceedings of the 2010 IEEE International Conference on Automation and Logistics*, 2010, pp. 379–384.

[24] P. Langley, J. E. Laird, and S. Rogers, “Cognitive architectures: Research issues and challenges,” *Cogn. Syst. Res.*, vol. 10, no. 2, pp. 141–160, 2009, doi: 10.1016/j.cogsys.2006.07.004.

[25] F. Zhuang *et al.*, “A Comprehensive Survey on Transfer Learning,” *Proc. IEEE*, vol. 109, no. 1, pp. 43–76, 2021, doi: 10.1109/JPROC.2020.3004555.

[26] C. Deng, X. Ji, C. Rainey, J. Zhang, and W. Lu, “Integrating Machine Learning with Human Knowledge,” *iScience*, vol. 23, no. 11, p. 101656, 2020, doi: 10.1016/j.isci.2020.101656.

[27] K. Lyytinen, J. V Nickerson, and J. L. King, “Metahuman systems = humans + machines that learn,” *J. Inf. Technol.*, p. 0268396220915917, May 2020, doi: 10.1177/0268396220915917.

[28] C. Emmanouilidis, E. Jantunen, and J. MacIntyre, "Flexible software for condition monitoring, incorporating novelty detection and diagnostics," *Comput. Ind.*, vol. 57, no. 6, pp. 516–527, 2006, doi: 10.1016/j.compind.2006.02.012.

[29] A. S. Pinto, "Novelty Detection for Semantic Place Categorization," MSc thesis, University of Porto, 2011.

[30] S. Budd, E. C. Robinson, and B. Kainz, "A survey on active learning and human-in-the-loop deep learning for medical image analysis", *Med. Image Anal.* vol. 71, 102062 (2021). https://doi.org/10.1016/j.media.2021.102062.

[31] N. Rubens, M. Elahi, M. Sugiyama, and D. Kaplan, "Active Learning in Recommender Systems," In: F. Ricci, L. Rokach, and B. Shapira (eds), *Recommender Systems Handbook*. Springer, Boston, MA. (2015), https://doi.org/10.1007/978-1-4899-7637-6_24.

[32] P. Pistofidis, C. Emmanouilidis, A. Papadopoulos, and P. N. Botsaris, "Management of linked knowledge in industrial maintenance," *Ind. Manag. Data Syst.*, vol. 116, no. 8, pp. 1741–1758, 2016, doi: 10.1108/IMDS-10-2015-0409.

[33] A. Kolus, R. Wells, and P. Neumann, "Production quality and human factors engineering: A systematic review and theoretical framework," *Appl. Ergon.*, vol. 73, no. May, pp. 55–89, 2018, doi: 10.1016/j.apergo.2018.05.010.

[34] M. Sony and S. Naik, "Industry 4.0 integration with socio-technical systems theory: A systematic review and proposed theoretical model," *Technol. Soc.*, vol. 61, no. April, p. 101248, 2020, doi: 10.1016/j.techsoc.2020.101248.

[35] G. R. Oldham and J. Richard Hackman, "Not what it was and not what it will be: The future of job design research," *J. Organ. Behav.*, vol. 31, no. 2–3, pp. 463–479, 2010, doi: 10.1002/job.678.

[36] "100 years of work design research: Looking back and looking forward," *J. Appl. Psychol.*, vol. 102, no. 3, pp. 403–420, 2017.

[37] F. P. Morgeson and S. E. Humphrey, "Job and team design: Toward a more integrative conceptualization of work design," *Res. Pers. Hum. Resour. Manag.*, vol. 27, no. July, pp. 39–91, 2008, doi: 10.1016/S0742-7301(08)27002-7.

[38] C. Spinuzzi, "The methodology of participatory design," *Tech. Commun.*, vol. 52, no. 2, 2005.

[39] E. B.-N. Sanders and P. J. Stappers, "Co-creation and the new landscapes of design," *CoDesign*, vol. 4, no. 1, 2008, doi: 10.1080/15710880701875068.

DOI: 10.1561/9781680838770.ch10

Chapter 10

A Review of Industrial Standards for AI in Manufacturing

By Eva Coscia, Rubén Alonso and John Soldatos

This chapter provides a review of industrial standards relating to Artificial Intelligence (AI) in manufacturing, including: (i) recommendations for human centric manufacturing systems; and (ii) technical standards for safety, security and data management.

10.1 Introduction

This chapter provides a review of industrial standards relating to Artificial Intelligence (AI) in manufacturing, including recommendations for human centric manufacturing systems and safety, security and data management related technical standards. The objective of this chapter is to highlight some of the most relevant standards, while detailing the possible considerations when designing and developing AI applications for the manufacturing industry such as those developed in the STAR[1] project.

The number of published standards, recommendations and reference architectures that can influence the development of AI solutions for the manufacturing sector is considerable, and as technology advances, new standards and recommendations are being worked on; therefore in this chapter we aspire only to provide some indications, either through links to standardization groups, or by providing a bit more details on a selection of the standards that we have found relevant for the development of AI systems in the manufacturing industry.

The chapter is divided as follows: first we present a review of the literature, citing various efforts to compile a list of standards. Then, we present several remarkable standards and of reference architectures. Finally, we present our conclusions and the need for further monitoring of the progress of various standards.

10.2 State of the Art Analysis

There have been several efforts in the literature to conduct an analysis of standards for the manufacturing sector. In general, due to the wide scope and the extensive collection of standards available, these analyses, like ours, focus on a specific part of the spectrum. Some of the most relevant analyses are detailed below.

Choi *et al.* [1], presented an analysis based on Factory Design and Improvement (FDI) process and the ISA-88 hierarchical model of manufacturing operations. In their document they focus on PPR (Product, Process, Resource) standards, categorizing standards related to Product data (e.g. DXF (Drawing Interchange Format), IGES (Initial Graphics Exchange Specification), VRML (Virtual Reality Modelling Language)), Process data (e.g OAGIS (Open Applications Group Integration Specification), ANSI/ISA-95) and Resource data (e.g. B2MML (Business to Manufacturing Markup Language), AP242) and their coverage on the FDI functional matrix.

Li *et al.* [2] reviewed several smart manufacturing standards and analyzed several industrial architectures. In particular, they focused on the standards developed by the following standard development organizations (SDOs):

- ISO/TC184 automation systems and integration, which develops standards related to information systems, control devices or data integration and interoperability.

[1] STAR – Safe and Trusted Human Centric Artificial Intelligence in Future Manufacturing Lines (https://star-ai.eu/).

- IEC/TC65 industrial-process measurement, control and automation., focused on activities that impact the integration of components, as well as different aspects of such systems, such as safety and security.
- ISO/IEC/JTC1 information technology, focused on ICT standards in different scopes, including security, multimedia or smart cards among others.

The National Institute of Standards and Technology (NIST) published in 2016 a landscape of standards focused especially on Smart Manufacturing Systems, which among other things details standards related to the different phases of product development, from design to end-of-life and recycling. This landscape covers modeling, data exchange, production system engineering and operation and maintenance standards, and identifies 8 priority areas in which standardization should advance: (i) Smart Manufacturing System reference model and reference architecture; (ii) Internet of Things reference architecture for manufacturing; (iii) Manufacturing service models; (iv) Machine to machine communication; (v) PLM (Product Lifecycle Management)/MES (Manufacturing Execution System)/ERP (Enterprise Resource Planning)/SCM (Supply Chain Management)/CRM (Customer Relationship Management) integration; (vi) Cloud manufacturing; (vii) Manufacturing sustainability; and (viii) Manufacturing cybersecurity.

W. Ziegler in [4] analyses the standardization landscape in the AI field, by focusing on the standards developed by five international and European SDOs (IEEE, ISO/IEC, ITU-T, ETSI, CEN-CENELEC) and two standards setting organizations (SSO): W3C and IRTF (Internet Research Task Force).

One of his main points is that even if the SDOs and SSOs are doing actions in the AI field, the standardisation activities are limited, and their number is low and does not increase at the same rate as developments and applications.

A new European player in standardisation activities and pursuing to increase the list of available standards on AI is the OASIS Open Europe Foundation (OOEF).[2] OOEF is the European sovereign affiliate organisation to the international nonprofit, OASIS Open, works to advance and support Europe's role in open source and open standards development. OOEF activities include: participation in collaborative projects supported by the EU and EU Member States, organization and participation in events for promoting the adoption of open source projects, engagement in European-specific activities to progress open source and open standards. The list of standards relevant for the Manufacturing domain and the AI technology covers Communication/messaging protocols (AMQP (Advanced

[2] https://www.oasis-open.eu/

Message Queue Protocol),[3] MQTT[4]), Cloud service management (TOSCA (Topology and Orchestration Specification for Cloud Applications)[5]), privacy management (PMRM[6]), security and production planning (PPS[7]), among other topics.

10.3 Standards Overview

As STAR is a project related to AI, safe and secure systems in the manufacturing domain, we have focused on reviewing standards from 3 major groups: (i) Technical, Management and Security Standards; (ii) Safety and Health Standards; and (iii) Other Relevant Standards. The list and categories were suggested by project partners (i.e.: technology providers, middleware developers or end-user, both SMEs and bigger firms), since these standards have implications in the activities they perform. The following sections discuss some of these standards and their impact on both the industry and on the suppliers of AI solutions for the industry.

10.3.1 Technical, Management and Security Standards

It is essential to begin by highlighting several of the families of ISO and IEC standards related to security, like the well-known ISO 27001, 27002 and 27701.

ISO/IEC 27001 [10] (Information technology – Security techniques – Information security management systems) specifies the requirements for implementing an information security management system (ISMS), and enables the assessment and treatment of security risks, and the implementation and management of security controls. This standard covers 3 main points, impacting the IT systems in the manufacturing industry: (i) the understanding and monitoring of security risks (including the knowledge of potential vulnerabilities, threats and the impact of them); (ii) design and implementation of security controls to reduce the risks; and (iii) the implementation of the process for the continuous management of the information security requirements.

ISO/IEC 27002 [11] (Information technology – Security techniques – Code of practice for information security controls), is applicable to organisations (public or

[3] https://www.oasis-open.org/committees/tc_home.php?wg_abbrev=amqp

[4] https://www.oasis-open.org/committees/tc_home.php?wg_abbrev=mqtt

[5] https://www.oasis-open.org/committees/tosca/

[6] http://docs.oasis-open.org/pmrm/PMRM/v1.0/cs02/PMRM-v1.0-cs02.html

[7] https://www.oasis-open.org/committees/tc_home.php?wg_abbrev=pps

private) of any type and size, including SMEs, and provide guidelines for implementation and management of security controls, allowing organisations to select controls for implementing ISO 27001 based ISMSs. The main objective behind this standard is to facilitate the selection of controls among well-known security controls and assists in the creation of guidelines for the management of the organization's ISMS. The standard covers issues related to data access, cryptography, asset management and information exchange, all of which are of importance when designing safe and secure systems for manufacturing.

ISO/IEC 27701 [5] (Security techniques – Extension to ISO/IEC 27001 and ISO/IEC 27002 for privacy information management – Requirements and guidelines) addresses the privacy and data protection perspective, and targets to enhance the Information Security Management System with additional privacy information management requirements. The standard allows Personally Identifiable Information Controllers and Personally Identifiable Information Processors to manage privacy controls and facilitate the organisations to implement efficient Privacy Information Management Systems, including policies and procedures for personal information management, including those needed to align the policies to privacy and Data protection regulations.

In addition to these privacy and security related standards, it is interesting to acknowledge the efforts of other technical committees that are publishing standards related to IA, security and secure data exchanges between devices and factory systems.

W3C XML security standards[8] are a family of neutral, open, vendor independent, and freely available standards, specifying security extension for the usage and interchange of XML data, which is one of the most used document encoding and data exchange formats. Among them, we can name: XML signature, for integrity, signer and message authentication; XML encryption for specifying the process of encrypting data and representing the resulting data in XML format; and XKMS for defining protocols for registering and distributing public keys, for use, for example, with XML signature and encryption.

One markup language used in the industry is AutomationML, this standard is covered in the IEC 62714 [12] (Engineering data exchange format for use in industrial automation systems engineering – Automation Markup Language). The purpose of this standard is to specify a data exchange format tailored to the needs of production system engineering and to enable the exchange of information, among heterogeneous engineering tools and, along the entire life cycle of production systems. AutomationML combines and adapts industry standards to

[8] https://www.w3.org/standards/xml/

reduce the data interchange and integration issues. The standards currently integrated in AutomationML are: CAEX for object topologies, hierarchies, properties, COLLADA for geometries and kinematics, PLCopen XML for discrete behavior of objects.

ETSI Cybersecurity and AI Standards[9] offer market-driven cyber security standardization solutions, recommendations and guidelines, and the improvement of the security of AI. These groups have a twofold objective: Understand and reduce cross-domain cybersecurity implications, and ensure that the artificial intelligence is secure. Network security, security of sensors and IoT devices, cybersecurity tools or machine-to-machine security are some of the topics covered by the ETSI CYBER standards. ETSI SAI is focusing on 3 main topics, all of them impacting the manufacturing domain: (i) Securing and protecting AI components from attacks; (ii) Mitigation against malicious and dangerous AI; and (iii) Using AI to improve security measures. Notable standards from these committees are: ETSI GR SAI 004: Securing Artificial Intelligence (SAI); Problem Statement, ETSI GR SAI 005: Securing Artificial Intelligence (SAI); Mitigation Strategy Report, or ETSI TR 103 787-1 CYBER; Cybersecurity for SMEs; Part 1: Cybersecurity Standardization Essentials.

ISO/IEC 18033 [13] (Information technology – Security techniques – Encryption algorithms) is a family of standards providing definitions, recommendations and guidelines for data encryption and confidentiality. This family of standard includes, so far, 5 different types of cyphers and encryption methods: (i) asymmetric ciphers; (ii) block ciphers; (iii) stream ciphers; (iv) Identity-based ciphers; (v) Homomorphic encryption. Both block ciphers and stream ciphers are in everyday use for transmitting information in industry and other sectors. Public key mechanisms are also being used for integrity and authenticity (e.g. certificates). Identity-based and Homomorphic systems are less widely used currently but are of interest for simplifying asymmetric encryption or for performing operations on encrypted stored data.

ISO/IEC 29100 [14] (Information technology – Security techniques – Privacy framework) establishes a framework for the protection of PII (biometric identifiers, names and surnames, location information, …) and establishes recommendations on how PII should be identified, how data should be controlled and how this data should be transmitted. This framework is applicable to various industries, including the manufacturing sector, and allows the organization to control security risks, comply with legal requirements, and reduce potential privacy breaches, which in turn can impact the organization's image.

[9] https://www.etsi.org/committee/cyber and https://www.etsi.org/committee/sai

ISO/IEC 27040 [15] (Information technology – Security techniques – Storage security) defines storage security terminology, details some scenarios related to the secure storage of data, and provides guidance on the security aspects associated with storage and storage technologies. Data storage and warehousing is a key issue in manufacturing information systems, and especially in the use of AI in the manufacturing industry. AI models require information to be trained and information to be used. This information is stored for decision making or for visualization. How, where and with what protection to store this data requires knowing the risks and implementing a secure storage approach.

10.3.2 Safety and Health Standards

Van Acker [6] in his dissertation on mental workload monitoring in the manufacturing industry, details several studies that show the relationship between fatigue and frustration and safety risks, loss of quality or performance, as well as detailing how changes in health (physical and mental) effects on workers lead to changes in burnout and job satisfaction. Ergonomics, the correct design of tasks and workplaces has a direct impact on the safety and security of workers. The incorporation of standards related to ergonomics and improving the quality of the worker's environment also impacts on the safety of innovative manufacturing environments. Among these standards, we can mention ISO 10075 [16] (Ergonomic principles related to mental workload). This standard defines terms related to mental workload, stress and strain, their consequences and the relationship between them. It also suggests methods to measure and assess the mental workload and provide requirements for measurement instruments, and provides guidance about the design of the workplace, equipment and activities.

Several other standards exist related to ergonomics and the design of computer and industrial systems. For example, ISO 9241 [17] (Ergonomics of Human System Interaction) , is a collection of standards that includes documentation and suggestions related to workplace ergonomics, visual displays, haptic and tactile interfaces, or general software ergonomics. Workspace design is also covered in ISO 6385 [18] (Ergonomics principles in the design of work systems), which, among various work environments, describes and provides guidelines for the design of work systems in production spaces, such as assembly line work. Human-centered design is also included in ISO/TR 16982 [19] (Ergonomics of human-system interaction – Usability methods supporting human-centred design). In the latter, human-centered usability methods are described and the advantages and disadvantages of these methods are presented. Lastly, ISO 26800 [20] (Ergonomics – General approach, principles and concepts) presents ideas applicable to the design of tasks or products, along with guidelines for tasks, products or even work areas to be efficient and safe.

Among those related to safe systems, we would like to mention two: (i) ISO 12100 [21] (Safety of machinery – General principles for design – Risk assessment and risk reduction); and (ii) ISO 45001 [22] (Occupational health and safety). The former specifies a risk analysis for machinery design. The second focuses on improving worker conditions and increasing employee safety by reducing workplace hazards. Both offer a framework for the control of risk factors, the mitigation of potential adverse hazards and the impact on the worker's physical and mental condition.

Last but not least, ISO/TS 15066 [23] (Robots and robotic devices – Collaborative robots) complements the ISO 10218 standard [24] (Safety Requirements for Industrial Robots) with focus on collaborative robots. Specifically, this TS describes 4 collective operation techniques:

- Safety-rated monitored stop
- Hand guiding
- Speed and separation monitoring
- Power and force limiting

Understanding this collaboration and knowing the risks is essential to avoid incidents resulting from robot-human contact. Maximum speed control, emergency stops, immediate contact stop and other technologies are some of the technologies applicable to industrial robots and especially to cobots.

10.3.3 Other Relevant Standards

Detailing all the relevant standards is beyond our scope, but there are several that are not specifically related to IA, safety, security or industrial environments, but are of interest from an industry point of view.

For example, ISO 9000 family [25] on Quality management, a well-known framework for demonstrating that products and services meet customer quality requirements, has some impact on IA systems for industry. In fact, there are several studies [7, 8] that are studying the use of IA for the management of incomplete, conflicting data or the classification of audit findings. Therefore, information from IA or the use of IA to improve decision making can be useful in improving quality management.

The commitment to responsibility and respect for society is covered in ISO 26000 [26] (Social responsibility), another non-technical standard not strictly related to IA in manufacturing, but which somehow takes into account company reputation, commitment and health. As mentioned earlier in the ergonomics standards, the health and safety of workers, especially in collaboration with machines, is a point of interest for the ecosystem of safe and secure AI systems.

In line with sustainability, mentioned above, we have ISO 14001 [27] (Environmental management systems). This standard, which like the previous one is not focused on manufacturing and AI, focuses on the introduction of environmental management systems for the improvement of sustainability and reducing potential conflicts related to the environment and even creating decent and healthy work environments [9].

Finally, it is worth mentioning a standard related to risk management in general. This standard is ISO 31000 [28] (Risk management), and is based on the evaluation and continuous optimization of the processes and how to manage risks. The standard, which targets the operational continuity of the business, takes into account, among other things, the safety of the outcomes or even the environmental reputation.

10.4 Industrial Architectures and Infrastructures

10.4.1 Reference Architecture Models

The advent of Industry 4.0 and smart manufacturing has given rise to the specifications of various architectures and reference models for developing digitally-enabled industrial systems, notably systems that leverage Cyber Physical Production Systems (CPPS). These models describe the structuring principles and the main building blocks of modern industrial systems. In most cases these reference models lead to industrial systems that fall in the realm of the Industrial Internet of Things (IIoT). The latter include data-intensive components such as components based on Big-Data analytics and Machine Learning (ML). As such their structuring principles are relevant to the development of AI systems. Following paragraphs review some of the most popular reference models for architecting Industry 4.0 systems including AI systems.

10.4.1.1 Reference Architecture Model Industrie 4.0 (RAMI 4.0)

RAMI4.0 provides structuring concepts and a vocabulary for understanding Industry 4.0 systems and their deployment. RAMI describes the structure and main elements of Industry 4.0 system by means of a 3D layered model (Figure 10.1). The three layers of the 3D model correspond to:

- The Architecture axis (Layers), which comprises six different layers indicating functionalities at different granularities of the system, from the asset to the business level.

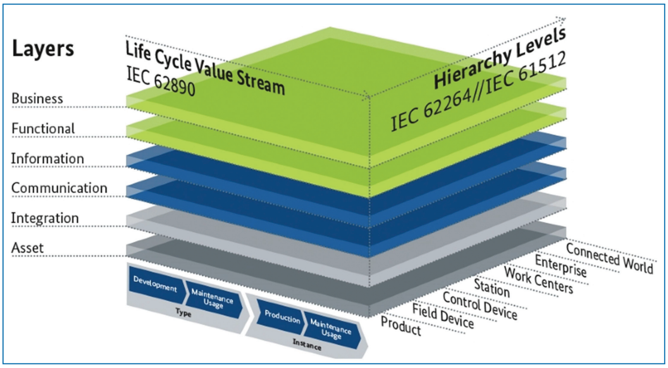


Figure 10.1. Reference Architecture Model Industry 4.0 (RAMI 4.0).

- The Process axis (Value Stream), which illustrates the stages of an asset's lifecycle, along with a corresponding value creation process based on IEC 62890.
- The Hierarchy axis (Hierarchy levels), which presents describe the breakdown structure of assembled components based on a taxonomy that starts from the product and goes up to the connected smart factory. The various levels are driven by the DIN EN 62264-1 and DIN EN 61512-1 standards.

The architecture layers of RAMI4.0 include:

- The Asset Layer, which describes physical systems and components (e.g., machines, motors, software applications, spare parts).
- The Integration Layer, which links the physical and digital/cyber worlds based on components like drivers and middleware.
- The Communication Layer, which deals with communications between the integration and information layers. It employs network protocols (e.g., TCP/IP, HTTP, FTP) over LAN and WAN networks, including wireless networks.
- The Information Layer, which provides (digital) information about sales, purchase orders, suppliers, locations etc. along with information on materials, machines and components that support the production.
- The Functional Layer, which comprises production rules, actions, processing, and system control.
- The Business Layer, which is associated with the business strategy, the business environment, and business goals of the enterprise, including promotions, offers, pricing models and cost analysis.

Process Value Streams

The Process axis deals with the lifecycle and processes of an object, which typically comprises a product, physical entities (e.g., machines and spare parts) or even virtual entities (e.g., documents and project plans). Every product needs to be updated, restructured, redesigned, or reformed for maintenance purposes. In this context, the process layer specifies Types and Instances as it main methods. A product in a development state is referred to as a "Type". Once moved to production, it becomes an "Instance". As illustrated in the RAMI4.0 cube, a "Type" is also subject to maintenance activities. A product returns to the "Type" state, whenever it is redesigned, or a new feature is being added to it.

Hierarchical Levels

The hierarchy levels of the corresponding axis are as follows: (i) Product, which abstracts the product that is manufactured in a factory; (ii) Field device, such as sensor and electronic devices that capture and/or control data from the field; (iii) Control device, which corresponds to the Operational Technology (OT) that manages input and output. Prominent examples are PLCs (Programmable Logic Controllers) and DCSs (Distributed Control Systems); (iv) Station, which enables operators to coordinate several processes and monitoring the results, by means of automation systems such as SCADA; (v) Work Center, which keeps track of manufacturing information and parameters that enable quality management; (vi) Enterprise, which comprises the are core business processes (e.g., production planning, production scheduling, marketing and sales, financial modules) that are usually managed through an ERP system; (vii) Connected World, which deals with the interlinking of all stakeholders as part of their supply chain interactions (including information sharing and exchange among them).

10.4.1.2 The Industrial Internet Consortium Reference Architecture (IIRA)

The IIRA specifies a common architecture framework for developing interoperable IoT systems for different vertical industries. It is an open, standards-based architecture, which has broad applicability. The latter makes is a vehicle for interoperability, mapping and practical deployment of IoT technologies, as well as standards development. To ensure its broad applicability, the IIRA is fairly generic, abstract and high-level. Hence, it can be used to drive the structuring principles of an IoT system, without however specifying its low-level implementation details. It is also a very good vehicle for communicating concepts and facilitating stakeholders' collaboration.

Based on the analysis of multiple use cases in different sector, the IIRA presents the structure of IoT systems from four viewpoints, namely business, usage, functional and implementation viewpoints. Among these four viewpoints, it's the functional viewpoint that specifies the functionalities of an IIoT system. To this end, the functional viewpoints specifies distinct functionalities in the form of the so-called

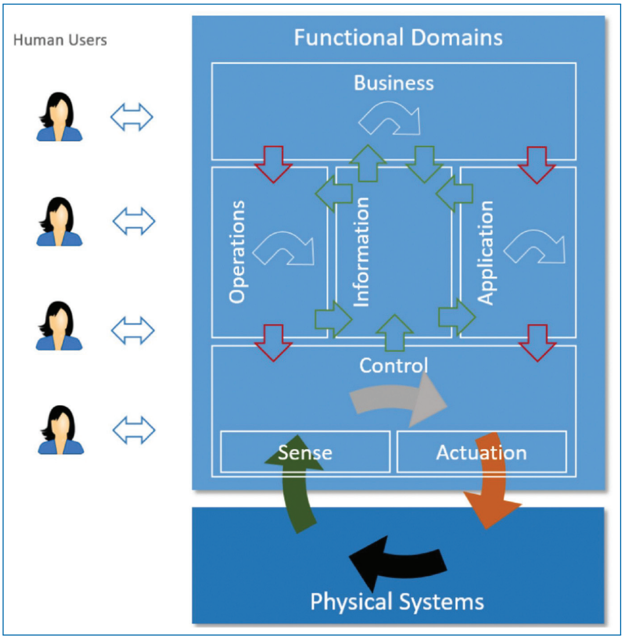


Figure 10.2. Functional domains in the IIRA.

“functional domains”. Functional domains (Figure 10.2) can be used to decompose an IoT systems in a set of important building blocks, which are applicable across different vertical domains and applications. As such functional domains are used to conceptualize concrete functional architectures. The IIRA decomposes a typical IoT/IIoT system into five functional domains, namely a control domain, an operations domain, an information domain, an application domain and a business domain as outlined in. The implementation viewpoint of the IIRA is based on a three-tier architecture, which follows the edge/cloud computing paradigm. It includes an edge, a platform and an enterprise tier.

10.4.1.3 The OpenFog Reference Architecture (RA)

The OpenFog Consortium was a consortium of high tech industrial enterprises companies and research/academic institutions, which are collaborating towards standardizing and promoting the fog computing paradigm. Since December 2018, the OpenFog Consortium and the IIC have joined forces.[10] Fog computing is directly associated with IoT, as it leverages fog nodes (i.e. essentially IoT devices) in order to enable reliable, low latency IoT applications. Fog computing alleviates

[10] https://www.iiconsortium.org/press-room/12-18-18.htm

the limitations and drawbacks of conventional cloud computing in various scenarios where low-latency and processing close to the field is required. The RA of the OpenFog consortium illustrates the structure of fog computing systems. It presents how fog nodes can be connected partially or fully to enhance the intelligence and operation of an IoT system. Moreover, it presents solutions about growing system wide intelligence away from low-level processing of raw data. The RA is described in terms of different views, including functional and deployment views. OpenFog compliant systems include some cross-cutting functionalities (i.e. functionalities that are applied across all layers of an IoT/OpenFog system). These cross-cutting functionalities are conveniently called "perspectives".

10.4.1.4 ISO/IEC CD 30141 Internet of Things Reference Architecture (IoT RA)

ISO/IEC 30141:2018[11] provides a standardized IoT Reference Architecture using a common vocabulary, reusable designs and industry best practices. It uses a top down approach, beginning with collecting the most important characteristics of IoT, abstracting those into a generic IoT Conceptual Model (CM). The latter has been derived based on a heuristic analysis of system characteristics that are common in most IoT systems (e.g., auto-configuration, discoverability, scalability, etc.). The CM describes typical IoT entities or actors, along with their relationships. The architecture is described by means of five complementary views i.e. functional, system, communications, information and usage.

10.4.1.5 BigData Value Reference Model

The BDV Reference Model provides the means for representing AI, ML, and BigData analytics pipelines [29]. It distinguishes between two different elements: (i) Elements that are at the core of the BDVA (Big Data Value Association); and (ii) Features that are developed in strong collaboration with related European activities. The model is structured into horizontal and vertical concerns:

- **Horizontal concerns** focus on the data processing chain, starting with data collection and ingestion, and extending to data visualisation. Horizontal concerns do not imply a layered architecture. For instance, visualisation may be applied directly to collected data without the need for intermediate functions like data processing and analytics.
- **Vertical concerns** address cross-cutting issues that apply to all horizontal functions and may include non-technical aspects.

[11] https://www.iso.org/standard/65695.html

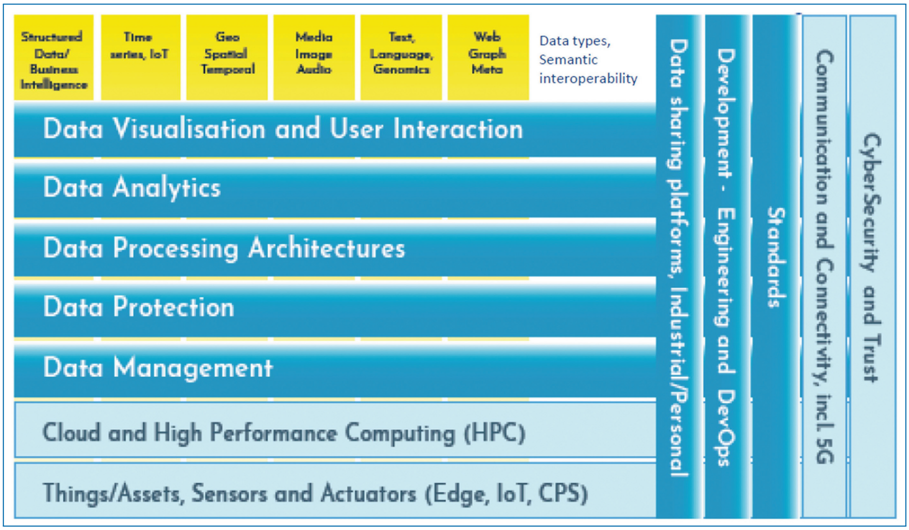

Structured Data/ Business Intelligence
Time series, IoT
Geo Spatial Temporal
Media Image Audio
Text, Language, Genomics
Web Graph Meta
Data types, Semantic interoperability
Data Visualisation and User Interaction
Data Analytics
Data Processing Architectures
Data Protection
Data Management
Cloud and High Performance Computing (HPC)
Things/Assets, Sensors and Actuators (Edge, IoT, CPS)
Data sharing platforms, Industrial/Personal
Development - Engineering and DevOps
Standards
Communication and Connectivity, incl. 5G
CyberSecurity and Trust

Figure 10.3. BDVA Reference Architecture Model for BigData Analytics and Machine Learning [29].

The BDV Reference Model (Figure 10.3) is compatible with reference architectures for AI, most notably to the ISO JTC1 WG9 Big Data Reference Architecture.[12]

10.4.2 Infrastructures for Industrial Systems

As far as infrastructures are concerned, we can mention GAIA-X[13] (a Federated Data Infrastructure for Europe). GAIA-X is an initiative launched by representatives from business, science and politics on a European level to create a proposal for the next generation of a European data infrastructure and thus enable EU companies to compete globally, exploiting data and services made available in an open digital and trusted ecosystem. GAIA-X connects centralised and decentralised infrastructures in order to turn them into a homogeneous, user-friendly system. The resulting federated form of data infrastructure strengthens the ability to both access and share data securely and confidently. Specifically for the Industry4.0 sector, GAIA-X infrastructure opens opportunities for development of new solutions for Smart Manufacturing, Supply Chain collaboration, Shared Production, Predictive Maintenance, Connected ShopFloor and more.

[12] https://www.iso.org/files/live/sites/isoorg/files/developing_standards/docs/en/big_data_report-jtc1.pdf

[13] https://www.data-infrastructure.eu/GAIAX/Navigation/EN/Home/home.html

10.5 Conclusions

The purpose of this chapter was twofold: on the one hand, to review some of the standards and architectures applicable to both, use cases and pilots, applicable to the manufacturing industry, that are being investigated in research and engineering projects, such as STAR. On the other hand, to provide a series of pointers and a brief overview of the current ecosystem of standards, so that those SMEs that want to implement safe and secure AI systems, either with or without middlewares such as STAR, can find some initial information.

Standards advance, are updated and new ones appear, that's why it is interesting to be aligned with the standards of suppliers and customers, and to perform periodic watch activities. For example, in the short-medium term there are several standards under development that may be of interest for the development of AI applications in the industry. Several of those included in *ISO/IEC JTC 1/SC 42 Artificial intelligence*,[14] such as *ISO/IEC AWI TR 5469 Artificial intelligence – Functional safety and AI systems*,[15] or *ISO/IEC DTR 24027 Information technology – Artificial Intelligence – Bias in AI systems and AI aided decision making*,[16] may have a high impact on the future AI for safe and secure systems.

Acknowledgement

This work has been carried out in the H2020 STAR project, which has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 956573.

References

[1] Choi, S., *et al.* "An analysis of technologies and standards for designing smart manufacturing systems." Journal of research of the national institute of standards and technology 121 (2016): 422–433.

[2] Li, Q., Tang, Q., Chan, I., Wei, H., Pu, Y., Jiang, H., … and Zhou, J. (2018). Smart manufacturing standardization: Architectures, reference models and standards framework. *Computers in Industry*, *101*, 91–106.

[14] https://www.iso.org/committee/6794475.html

[15] https://www.iso.org/standard/81283.html

[16] https://www.iso.org/standard/77607.html

[3] Lu, Y., Morris, K. C., and Frechette, S. (2016). Current standards landscape for smart manufacturing systems. *National Institute of Standards and Technology, NISTIR, 8107*, 39.

[4] Ziegler, Wolfgang. "A Landscape Analysis of Standardisation in the Field of Artificial Intelligence." *Journal of ICT Standardization* (2020): 151–184.

[5] ISO/IEC 27701 Security techniques – Extension to ISO/IEC 27001 and ISO/IEC 27002 for privacy information management – Requirements and guidelines, International Organization for Standardization, Geneva, Switzerland. https://www.iso.org/standard/71670.html (Accessed: June 2021)

[6] Van Acker, B. (2020). *Mental Workload Monitoring in the Manufacturing Industry: Conceptualisation, Operationalisation and Implementation* (Doctoral dissertation, Ghent University).

[7] Neves, J., Fernandes, A., Gomes, G., Neves, M., Abelha, A., and Vicente, H. (2015, April). International Standard ISO 9001 an Artificial Intelligence View. In *ICEIS (1)* (pp. 421–428).

[8] Corpuz, R. S. A. (2019). Implementation of artificial neural network using scaled conjugate gradient in ISO 9001: 2015 audit findings classification. *Planning*, *5*(2), 5–3.

[9] Anwar, C., Purwanto, A., Abidin, R. Z., Prabowo, R. F., Rani, C. P., Saefulah, K. F., and Sulistyo, A. B. (2020). ISO 9001: 2015, ISO 14001: 2015, ISO 45001: 2018 AND ISO 22000: 2018: Which are the most affected manufacturing performance?. *Journal of Critical Reviews*, *7*(19), 2311–2330.

[10] ISO/IEC 27001 Information security management, International Organization for Standardization, Geneva, Switzerland. https://www.iso.org/isoiec-27001-information-security.html (Accessed: June 2021).

[11] ISO/IEC 27002 Information technology – Security techniques – Code of practice for information security controls, International Organization for Standardization, Geneva, Switzerland. https://www.iso.org/standard/54533.html (Accessed: June 2021).

[12] IEC 62714 Engineering data exchange format for use in industrial automation systems engineering – Automation Markup Language, International Electrotechnical Commission. https://webstore.iec.ch/publication/32339 (Accessed: June 2021).

[13] ISO/IEC 18033 Information technology – Security techniques – Encryption algorithms, International Organization for Standardization, Geneva, Switzerland. https://www.iso.org/standard/76156.html (Accessed: June 2021).

[14] ISO/IEC 29100 Information technology – Security techniques – Privacy framework. International Organization for Standardization, Geneva,

Switzerland. https://www.iso.org/standard/45123.html (Accessed: June 2021).

[15] ISO/IEC 27040 Information technology – Security techniques – Storage security, International Organization for Standardization, Geneva, Switzerland. https://www.iso.org/standard/44404.html (Accessed: June 2021).

[16] ISO 10075 Ergonomic principles related to mental workload, International Organization for Standardization, Geneva, Switzerland. https://www.iso.org/standard/66900.html (Accessed: June 2021).

[17] ISO 9241 Ergonomics of human-system interaction, International Organization for Standardization, Geneva, Switzerland. https://www.iso.org/standard/77520.html (Accessed: June 2021).

[18] ISO 6385 Ergonomics principles in the design of work systems, International Organization for Standardization, Geneva, Switzerland. https://www.iso.org/standard/63785.html (Accessed: June 2021).

[19] ISO/TR 16982 Ergonomics of human-system interaction – Usability methods supporting human-centred design, International Organization for Standardization, Geneva, Switzerland. https://www.iso.org/standard/31176.html (Accessed: June 2021).

[20] ISO 26800 Ergonomics – General approach, principles and concepts, International Organization for Standardization, Geneva, Switzerland. https://www.iso.org/standard/42885.html (Accessed: June 2021).

[21] ISO 12100 Safety of machinery – General principles for design – Risk assessment and risk reduction, International Organization for Standardization, Geneva, Switzerland. https://www.iso.org/standard/51528.html (Accessed: June 2021).

[22] ISO 45001 Occupational health and safety, International Organization for Standardization, Geneva, Switzerland. https://www.iso.org/iso-45001-occupational-health-and-safety.html (Accessed: June 2021).

[23] ISO/TS 15066 Robots and robotic devices – Collaborative robots, International Organization for Standardization, Geneva, Switzerland. https://www.iso.org/standard/62996.html (Accessed: June 2021).

[24] ISO 10218 Robots and robotic devices – Safety requirements for industrial robots, International Organization for Standardization, Geneva, Switzerland. https://www.iso.org/standard/51330.html (Accessed: June 2021).

[25] ISO 9000 family – Quality management, International Organization for Standardization, Geneva, Switzerland. https://www.iso.org/iso-9001-quality-management.html (Accessed: June 2021).

[26] ISO 26000 Social responsibility, International Organization for Standardization, Geneva, Switzerland. https://www.iso.org/iso-26000-social-responsibility.html (Accessed: June 2021).

[27] ISO 14000 family – Environmental management, International Organization for Standardization, Geneva, Switzerland. https://www.iso.org/iso-14001-environmental-management.html (Accessed: June 2021).

[28] ISO 31000 Risk management, International Organization for Standardization, Geneva, Switzerland. https://www.iso.org/iso-31000-risk-management.html (Accessed: June 2021).

[29] BDV SRIA European Big Data Value Strategic Research and Innovation Agenda, BDVA, Version 4.0, October 2017, available: https://bdva.eu/sites/default/files/BDVA_SRIA_v4_Ed1.1.pdf

DOI: 10.1561/9781680838770.ch11

Chapter 11

AI That Works: The Symbiosis of Functionals & Non-Functionals as Main Success Factor

By Arthur van der Wees, Anna Ida Hudig and Celine Prins

With any emerging technology, or combination of existing technologies, one tends to focus on the technology itself. However, the technology should not be the focal point as it in itself is not the solution. This also goes for Artificial Intelligence and the promising functionalities and capabilities it can or otherwise promises to bring, enable, facilitate and augment. For instance in the vast supply chains, manufacturing, logistics, maintenance and related Industry 5.0 domains. This chapter will present notions and guidance to make AI work; not just function but also to have it prepared by design with embedded non-functionals for when things may go wrong and other risks it may encounter or cause. All this for AI to help making 'it' work. This tiered approach provides value propositions that effectively address societal challenges, for which relevant AI functionalities in symbiosis with risk-based non-functionalities can be designed, deployed and continuously improved.

In the Industry 5.0 domain this approach is aimed to result into valuable and feasible, human-centric, secure, safe, sustainable and otherwise trusted and trustworthy AI-supported intelligent ecosystems. With that, the symbiotic, dynamic equation of both functionals and non-functionals is one of the main success factors for future-proof Industry 5.0 and related value creation.

11.1 Introduction

11.1.1 How to Make it Work?

Where this chapter mainly aims to provide guidance in making Artificial Intelligence work, in order to get there it is important to first understand how to make 'it' work – and what 'it' actually means –.

11.1.2 Everything is Connected

Alexander Von Humboldt, the 18th-century scientist, naturalist and explorer, world famous in his time, was one of the first to recognize and explain the fundamental functions of the mountains, rivers and rain forests for the ecosystem and climate, claiming that the world is a single intertwined and interconnected organism.

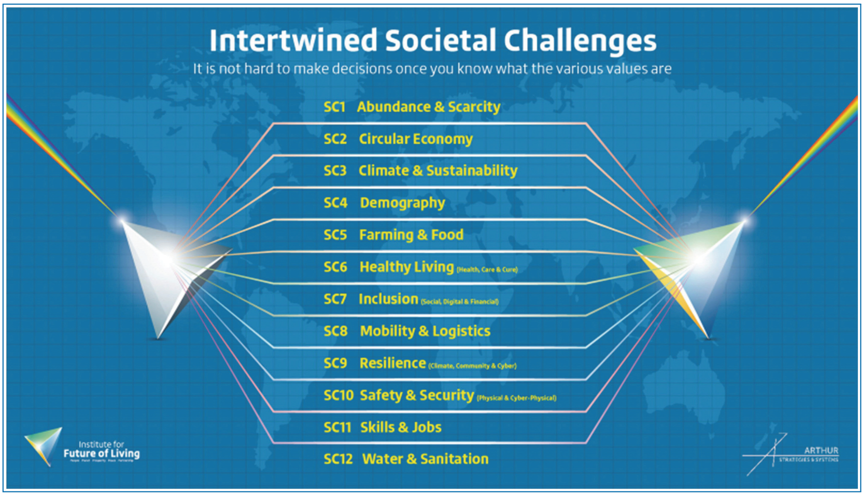
Everything is connected; everything is part of the system, he basically acclaimed [1]. This is the concept of nature as we know it today. According to Von Humboldt, everything, to the smallest creature, has its role and together makes the whole, in which humankind is just one small part of the holistic puzzle.

Integrated ecosystems sustain life and provide us with an amazing habitat. People and the ecosystems we live in, in this Digital Age, have great capabilities to improve and sustain the quality of life for all. If we interact and leave no one behind. As we face and urgently need to deal with many societal challenges, we need a climate for change. For these, Artificial Intelligence (AI) and other or related knowledge, processes, technologies, human intelligence and experience may be an excellent enabler and facilitator.

11.2 Where to Start?

11.2.1 Societal Challenges

Where this chapter is not about covering and discussing societal challenges in general, one needs valued use cases that address any or multiple of those challenges. This, as technology in itself is not a use case. It never is. The way forward is to have



Intertwined Societal Challenges

It is not hard to make decisions once you know what the various values are

SC1 Abundance & Scarcity
SC2 Circular Economy
SC3 Climate & Sustainability
SC4 Demography
SC5 Farming & Food
SC6 Healthy Living (Health, Care & Cure)
SC7 Inclusion (Social, Digital & Financial)
SC8 Mobility & Logistics
SC9 Resilience (Climate, Community & Cyber)
SC10 Safety & Security (Physical & Cyber-Physical)
SC11 Skills & Jobs
SC12 Water & Sanitation

Figure 11.1. Intertwined societal challenges.

an use case-driven, people-centric, stakeholders-centric, persona-centric, societal-centric, data-centric, sustainable, technology-agnostic and accountable approach.

But, where to start? What can an entrepreneur, company, sector, community or other groups in society and economy do to create overall positive, green, digital and resilient impact while also having a viable and economically sustainable value model, with related business models and (financial and other) feasibility models to get things both started, going, trusted, growing, scaling, resilient and future-proof? Having a big vision and focusing on the horizon is important, but having a clear starting point is one of the main prerequisite success factors.

With that in mind, it is recommended to start with identifying and establishing the particular challenge(s) one would like to focus on, for instance by using the 12 Societal Challenges for Future of Living [2], as visualised below in Figure 11.1. These are in line with both the vision of the European Commission as well as the United Nations' Sustainable Development Goals (SDGs) [3]. These Societal Challenges are obviously intertwined and interconnected.

When analysing the various Societal Challenges that relate to the vast supply chains, manufacturing, logistics, maintenance and related Industry 5.0 domains [4] in combination with digital ecosystems with certain AI capabilities or potential anywhere upstream, midstream or downstream in the Industry 5.0 ecosystems, at least the following are notable to be considered:

SC1: Abundance & Scarcity
SC2: Circular Economy
SC3: Climate & Sustainability
SC4: Demography

SC7: Inclusion
SC8: Mobility & Logistics
SC9: Resilience (Climate, Community & Cyber)
SC10: Safety & Security
SC11: Skills & Jobs

As two examples, the next paragraphs will briefly dive into the SC4: Demography, and SC11: Skills & Jobs, and where and why AI in Industry 5.0 context may be valuable, appreciated and even necessary.

It will also demonstrate that each Societal Challenge is both complex in itself as well as intertwined with the other Societal Challenges, where addressing one also will result in addressing or otherwise impacting parts of others.

11.2.1.1 H2M & M2H cooperation

When focusing on the Societal Challenge of Demography, three questions that comes to mind are (a) how to deal with an expected decrease of population in multiple parts of Europe, (b) what will the various combinations and interfaces between humans and machines, and vice versa, look like (H2M, M2H, H2M2M and the like), and (c) will we see social prosperity or social disruption?

Within the European Union, there is a decline in working-age population. It's expected to reduce by 13.5 million (or 4%) by 2030 compared to 2018 [5]. This, as the EU population size will shrink by 5% between 2019 and 2070, to 424 million inhabitants [6]. Furthermore, the EU's demographic ratio between people above 65 years old and those aged 20-64 is expected to increase from a one to four ratio 2010, to a one to (less than) two in 2070 [7]. Additionally, the development of shorter working weeks could cause a 2% reduction in labour supply [8].

So, more productivity and efficiency is expected from less [9]. The developments will affect the per capita gross domestic product (GDP) but also welfare and the quality of life. Just keeping the current status quo in place will be a huge challenge. According to a Working Paper of the Organisation for Economic Cooperation and Development (OECD) on ageing and productivity growth, in many OECD regions the actual growth rates recorded have been lower than productivity growth required to maintain per capita GDP levels in recent years. One reason for this is that ageing also has a direct negative impact on productivity growth, with the effect being concentrated in urban areas [10].

Combining and deploying innovative processes, data and technologies to augment the capabilities of people, industry, supply side and demand side can be helpful mechanisms to compensate this expects decrease in productivity and levels of welfare and quality of life.

The above does not only demonstrate that are huge potential and markets for AI, intelligent systems, cognitive computing, robotic process automation, distributed intelligence, autonomous systems, cobots, and other or related knowledge, processes, technologies and experience. It also demonstrates that there is a need for AI- and other technology-supported H2M, M2H, H2M2M and other interaction, communication and cooperation to help address the current and upcoming challenges, avoid social disruption, and improve social prosperity.

11.2.1.2 Evolution or revolution?

When focusing on the Societal Challenge of Skills & Jobs, three questions that come to mind are (i) how will the future of work change the industrial sector, and the looks of our urban and rural societies, (ii) how to keep the veins of trade and human values running through our communities, and (iii) will technology displace more jobs in 10 years than it creates, or vice versa?

According to the OECD, 65% of the kids in schools today will have jobs that haven't been invented yet [11]. This indicates that we apparently are not yet sure what the future will look like, but that we do for sure acknowledge society will look very differently in a decade. The World Economic Forum (WEF) points out that among the top 10 most essential skills of the near future are: analytical thinking, empathy, creativity, reasoning, complex problem-solving, self-management, and technology development and use [12]. Clearly, this list resembles a more intertwined combination of both the right part of the brain with the left part, than currently commonly seems the case.

Intelligent supply chains, rapid innovation production, integrated logistics, prognostic health monitoring, predictive maintenance and other Industry 5.0 domains have the capabilities to address Societal Challenges and improve productivity, safety, security, sustainability and other efficiencies [13]. New concepts, models and processes supported with AI and other digital capabilities are not a nice to have; they are a need to have. For sure it will also both support and augment the workforce, yet it will also challenge and change it, in an evolutionary or revolutionary way.

However, Societal Challenges and related SDGs are challenging and complex problem sets. There is no one solution. There is no one entrepreneur, no one corporation, nor one other group with the answer. There is no one technical fixture. Nor will there be one AI fixture. This is all about working together. Each challenge requires diverse teams and capabilities. Nothing less. This is all about walking and achieving outcomes.

11.2.2 People, Process & Technology

The current real-life world in this Digital Age seems complex than ever. It is and will be more and more the symbiosis of physical, physical-cyber, cyber and cyber-physical worlds, with ever-increasing capabilities and possibilities. Therefore,

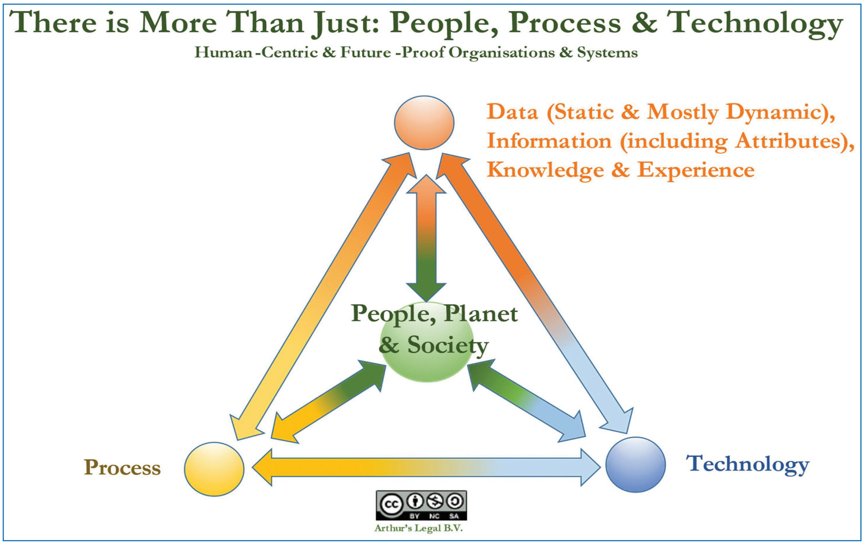

There is More Than Just: People, Process & Technology
Human-Centric & Future-Proof Organisations & Systems
Data (Static & Mostly Dynamic),
Information (including Attributes),
Knowledge & Experience
People, Planet
& Society
Process
Technology
BY NC SA
Arthur's Legal B.V.

Figure 11.2. Tetrahedron: people, process, technology & data.

it is important to identify, map and plot one of the all-present common denominator in this Digital Age.

Also for the European Commission, from the digital perspective, the common denominator and the main priority is: data [14]. The data dimension is the dynamic and all-present dimension that is relevant everywhere in this Digital Age. It offers huge opportunities, benefits and gains.

Nonetheless, in the generic structure of 'people, process, technology', the component 'data' is generally still overseen. Yet in the AI domain it is obviously one of the main ingredients and enablers. Data, structured data (being information), combined information (being knowledge), and used data, information and knowledge (being experience) in all its forms and categories and with all its various values, will generally run through all the activities and ecosystems, from end to end. The time has come that we all move away from just thinking in the traditional, and long-outdated, mode of 'people, process and technology'.

For addressing any Societal Challenges, one will need a data strategy, which we always need to take it in any equation, visualised below in Figure 11.2.

11.3 Make it Work

11.3.1 Make it Function

Technology changes the world at a fast pace. Thirty years ago the internet became publicly available through the World Wide Web. It was not designed. It just

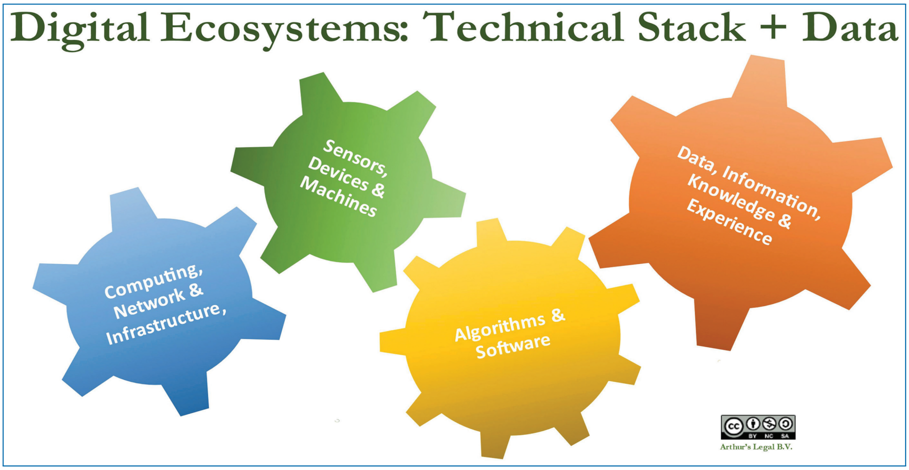

Digital Ecosystems: Technical Stack + Data
Sensors, Devices & Machines
Data, Information, Knowledge & Experience
Computing, Network & Infrastructure,
Algorithms & Software
BY NC SA
Arthur's Legal B.V.

Figure 11.3. Digital ecosystems are interconnected vessels.

happened and evolved. Meanwhile, it has fundamentally changed the world as we then knew it. Today we see more than 50% of the world's population have the ability to access and use these digital technologies and networks. And the number continues to increase, every day.

Societies and individuals can benefit in all manner of ways from access to knowledge, people and organizations on a local and global level. More than that, digital has become a must-have, for people, society and the economy. Indeed, digital technology and networks foster innovation. Digital platforms, AI, robotics, edge computing and the internet of things (IoT) are further expediting this process by connecting, inter-connecting respectively hyper-connecting individuals, organizations, communities, societies and data, with tens of billions of objects and entities.

All these technical capabilities and related digital ecosystems generally comprise of a technical stack that to some extent can be visualized as set forth below in Figure 11.3. These are made up of some combination of the various forms of data together with software-enabled algorithms that have sufficient computing power either centralized, decentralized or distributed on the Edge or in IoT devices, and interfaces, connectivity and infrastructure where necessary.

So, with the clear mission to address Societal Challenges one has done the initial preparations to make it work. When furthermore preparing the relevant kitchen tools, cooking ingredients, basic cooking skills and a plan what to cook, one can come up with the technical functions, and the functional specification, technical requirements, the technical specification, and thereafter the actual development and engineering. Right after, it is time to demonstrate it functions, and one is all set. Right? We all know how difficult it already is to even come to that point.

11.3.2 Does It Work? Or Does It Just Function?

However; does it actually work? Or does it just function? What if it does not function?

AI technology is an inherent component of Industry 5.0. However; even if the technology itself may be at the right technical readiness level, the readiness of a technology on itself, that is, whether it has been proven to be well-functioning in an operational environment, does not guarantee its success [15]. Studies show that adding AI to a technology or process could strengthen its capacity to reach the envisioned outcome, yet it will just as well amplify the risk for negative impact. Digital technologies and intelligent networks are not immune to error, evil, incidents or other risk. These are also not immune to incidental, incremental or disruptive change, either caused by internal or external factors. The many 'What-If' scenarios are generally not considered sufficiently, and not re-run after in a consistent and on continuous basis.

Making it work, implies having both the functionals as well as non-functionals included, by design and by default, and taken into consideration – and addressing those – end-to-end; both upstream, midstream and downstream, in the Von Humboldt spirit.

Although new and seemingly burdensome for some, it will for sure be beneficial in order to truly make it work, with AI in the equation. Before one notices, it will become second nature. The 'it', in 'make it work' is not AI or other technological functionalities or capabilities; it is a valued use case that addresses Societal Challenges of any kind.

11.3.3 Risk in Cyber-Physical and Other Digital Ecosystems

11.3.3.1 Risk in the digital age

Where this chapter is not aimed to give a full overview and perspectives of risk, risk mitigation and risk management, it is important not to see risk as something necessarily negative. It is an integral part of the equation and with that an enabler and facilitator of anything that works in a trusted, trustworthy and accountable way. It gives essential and valuable insights in what may happen or may go wrong, what people or society like or fear, et cetera. For sure, in the AI or AI-supported domain that is an essential success factor.

The magnitude of risks, determined by the probability as well as the impact thereof, is very much context and application dependent. To prepare for and mitigate the potential harm, to embed preparedness for foreseen and unforeseen situations, and to make it resilient and future-proof, it is necessary that AI systems are designed and deployed guided by trust principles. These non-functionals are principles that consistently preserve trust, trustworthiness and engagement of all relevant stakeholders. Examples of such principles are security, safety, privacy, transparency, auditability, sustainability and robustness. There are several hundred of trust principles. These can be found in best practices, guidelines, white papers, standards, regulations but also in common practice and nature.

Two major challenges in the AI design and deployment are (1) to map the relevant risks accurately and comprehensively throughout the system's entire lifecycle, and (2) to incorporate non-functionals by design.

11.3.3.2 Risk segmentation; creating insights & oversight

Risk is not a four-letter word, and – even in the AI context – deserves its own series of books. It is at least useful to segment the various AI-related dimensions of this Digital Age in order to get some relevant oversight and insight. For purposes of this book in general and this chapter in particular that argues for a holistic, end-to-end ecosystem approach, similar to the notions of Von Humboldt, the initial segmentation however done in four (4) segments as set forth below:

A. **Non-connected**, which is a stand-alone device, tool, machine, appliance or application that does not have connectors or connectivity that can connect to the internet or other external network or resources.

B. **Connected**, where a device, tool, machine, appliance, application or system may be connected to, via the internet, a centralised databases, cloud infrastructure and other centralised systems;

C. **Inter-connected**, where several edge devices, tools, machines, appliances, applications or systems are connected with each other, either via orchestrated, federated systems, and;

D. **Hyper-connected**, where numerous far edge and other IoT devices, tools, machines, appliances, applications or systems are directly connected with each other via distributed (computing and related) ecosystems of ecosystems.

For each of these segments, various value cases, business models, feasibility models and therefore use cases can be identified and created in the AI-supported Industry 5.0 domain. Each segment has its own values, benefits, efficiencies, inefficiencies, et cetera.

The segmentation set above obviously is not the only one possible. Various other segmentations are relevant to consider as well, such as for instance real-time, near-real-time or not. This segmentation may be relevant when near-real-time autonomous 3D printing is considered, or real-time prognostic health monitoring or related integrated logistics support are relevant. Other segmentations that can be considered are single-vendor, multi-vendor, OEM, public, private, public-private, et cetera.

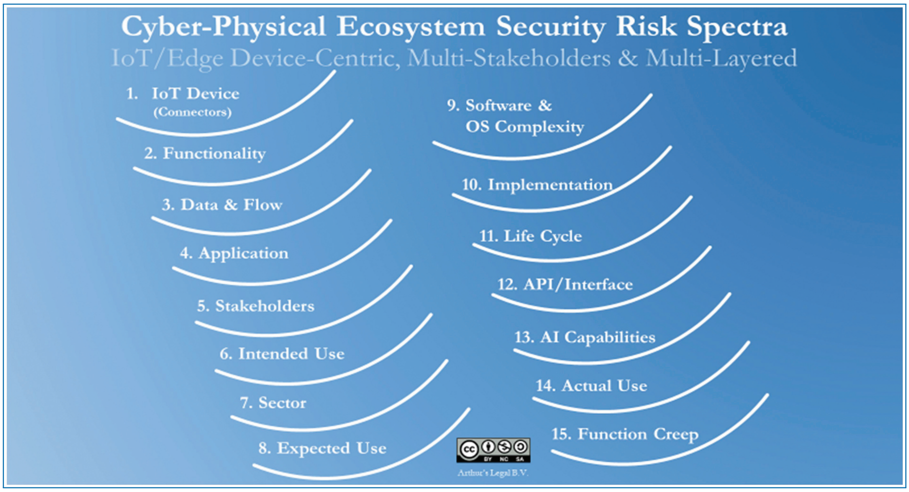


Figure 11.4. Cyber-physical ecosystem security risk spectra.

11.3.3.3 Risk classification spectra: a multi-layered approach

When going back to the above-mentioned segment, Hyper-Connected devices, and taking a risk-perspective to those, a methodology to do high-level quality risk classification is to have a multi-layered approach and do such risk classification per spectrum, starting with the risk classification of the connectors and connectivity of the IoT device itself. Even though AI capabilities may not yet be in the equation, it is essential to understand the various risks that are embedded in or could arise from such a IoT device. Subsequently, other risk spectra should be considered and risk classified, as visualised below in Figure 11.4.

Especially more downstream there may be risk spectra that may not be relevant; however, if such spectrum may become relevant later in the life cycle of the IoT device it is recommendable to keep it in and already do the spectrum risk classification. In general, three categories of main risk levels are used: low, medium and high. Based on the outcome of (i) a risk classification for each spectrum, and (ii) the interim outcome of the various risk classifications up to Spectrum 13 (AI Capabilities), the baseline risk classification can be established.

Based on that baseline, the AI Capabilities risk classification can be done, and the subsequent risk spectra; the holistic perspective constitutes the Combined Risk Classification, on which one can consider and organise technical & organisational security, safety, privacy and related technical and organisational measures.

Any technical and organisational measures taken or to be taken can include, cause or otherwise trigger risk by itself or as a trigger consequence. It is therefore recommended to double-loop the particular set of measures, for once to initially assess if and to what extent these may have a detrimental impact, for which in the subsequent section 'Double Looped S.I.M.' will provide with a practical and proven methodology.

As per the dynamics of IoT – and even more so AI-supported IoT and IoT ecosystems –, any of the risk classification spectra can be expected to trigger, change or otherwise show relevant dynamics, such as (A) technical or other threats and vulnerabilities, (B) actors and other stakeholders anomalies, updates or upgrades in code, datasets or attributes, or (C) changes in regulatory standards, policies or other relevant best practices, it is recommended to double-loop as well, including those spectra that are or may be related or otherwise are (inter)depended on the particular spectra. Therefore, it is recommended to continuously monitor the risks, and where necessary or otherwise double-loop thereafter to keep the security measures up to date and resilient.

In any case, the segments, whether non-connected, connected, inter-connected or hyper-connected, that have AI capabilities of any kind, are for sure game changing, where non-functional and functional requirements have to be addressed together. The winner will be the one who understands fully the societal challenges at hand and related sectoral requirements.

11.3.4 Good-Case, Bad-Case & Worst-Case Scenarios

11.3.4.1 Dynamic scenarios; identify, structure, act & double-loop

The world is not perfect. Nothing is. Nature and its dynamics are used to that, and so are humans – even though this is forgotten once in a while. Every professional and every other person has a lot of individual capabilities to assess risks, including probability, potential impact and severity thereof.

However, with an adequate number of individuals coming from diverse groups of people with diverse backgrounds, knowledge, skills and expertise, one can do even more comprehensive risk assessments. Multiple and various brain power leads to more perspectives, angles, understanding about interdependencies and other insights. This is very necessary as well, as the (i) ever-changing and ever-evolving systems and attack surfaces, and (ii) assessing risk and making informed decisions to choose, implement and keeping up to date the right set of technical and organisational measures, become more and more complex as well.

Risk is generally linked to accidents. These can have a low probability, but when these happen it can result in high impact and can even trigger multiple severe external consequences. In his well-known book *Normal Accidents*, Perrow uses the term 'normal accidents' in part as a synonym for 'inevitable accidents.' This categorization is based on a combination of features of such systems: interactive complexity and tight coupling. Normal accidents in a particular system may be common or

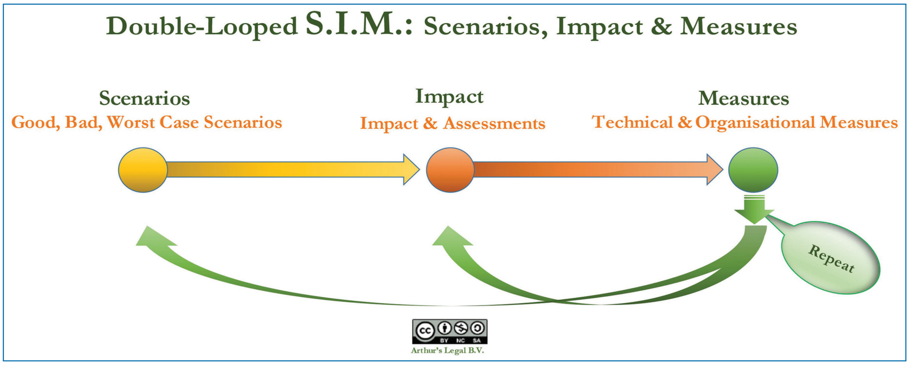

Double-Looped S.I.M.: Scenarios, Impact & Measures
Scenarios
Good, Bad, Worst Case Scenarios
Impact
Impact & Assessments
Measures
Technical & Organisational Measures
Repeat
BY NC SA
Arthur's Legal B.V.

Figure 11.5. Double-Looped S.I.M.

rare, but the system's characteristics make it inherently vulnerable to such accidents, hence their description as 'normal' [16].

However, risk is not just linked to accidents. It can also be a consequence of action, inaction, error, omission, ignorance, stupidity, Recklessness, intention, malicious or unintended but foreseeable consequence.

11.3.4.2 Double-looped scenario plotting

To consider risk in order to take technical and organisational measures by design, the scenario plotting methodology of Double-Looped S.I.M. can be used, preferably together with diverse groups of people with diverse expertise, and at different moments in time and different times during the day. S.I.M. means: Scenario, Impact, Measures. The double-looping refers to the notion that any measure in itself can be a vulnerability and can even increase risk or create new risk and related detrimental impact. So, every measure deserves its own S.I.M. cycle. The Double-Looped S.I.M. can be visualised as set forth below in Figure 11.5.

One will probably find numerous scenarios one would not immediately think about, as generally the worst-case scenarios get the most attention even though the probability may be close to 0%. Good case scenarios generally are forgotten, although they may cause (intially unforeseen) impact and negative consquences as well. With the nowadays familiar race to try to be the first in a market, the risks of AI and its functions and applications that have been designed, seemingly 'for good', may have severe negative societal, safety, security, personal, economical, ecological and other risks, impact and consequences. The ones that create AI are humans and generally working with a certain focus and under certain pressure (including by its investors, grant providers and others), while not considering or allowed to consider other perspectives. Furthermore, new or emerging technology tends to be overconfident. In the case of AI capabilities, even the AI may be overconfident itself [17].

Other scenario examples can – and should – be taken from real life accidents and other lessons learned. Such as the story of Wanda Holbrook, a maintenance technician and who was killed by an unexpectedly moving robot at a car parts manufacturer site [18]. Unfortunately, the non-functionals – in this case personal safety and security – was not taken into the symbiotic equation, and the total imbalance became awfully clear.

11.3.4.3 Balancing out functionals & non-functionals

The question 'what happens if things go wrong?' is not one most designers, developers or marketeers wish to ask themselves. Even more, in the AI domain it is expected that incidents will have an even more severe impact than in the digital domains without AI capabilities. These notions also go for any emerging or relatively new technical capabilities; not only for the AI domain. Good and extensive scenario plotting and mapping are prerequisite, also from ethical and accountability perspectives.

The appropriate balance between functionalities and benefits on the one hand, and non-functionals and impact-mitigation on the other hand, with appropriate security and other prevention-, risk- and impact-based measures, metrics and measurements in place will need to be found per context, and meanwhile monitored and challenged continuously. It will increase transparency, reduce unpleasant surprises in the Digital Age, and most of all increase trust and trustworthiness. Making it work, including the appropriate functionals, non-functionals and related accountability, is complex but that is where the true huge potential is, for all, and the future of mankind and our planet.

11.3.5 Human-Centric Co-Creation Cycle: Success by Design

11.3.5.1 Cat & Mouse

Same as in cat and mouse games, malicious actors immediately change and improve their ways as soon as they are countered. In AI but also any other technology or digital ecosystem, the eternal cat and mouse game will continue, increase and expedite. 'AI for Good' can easily be converted into 'AI for Malicious', and vice versa. Therefore, future networks will indeed be smarter and safer, whilst at the same time those networks will be more vulnerable. This race will not be a sprint; it will be a permanent marathon with an unknown number of sprints. In the Digital Age, these eternal games will continue, increase in dynamics and speed, and otherwise also expand exponentially.

In order to aim to identify risk and avoid that impact is mitigated or contained and vulnerabilities are not misused, the first focus should be on trying to avoid that there are risks and vulnerabilities in the first place, preferably by design and in a continuous manner. This is also not a task or responsibility of one person, one department or one organisation. No one can do this alone. For this, the Human-Centric Co-Creation Cycle methodology was developed, validated and deployed worldwide.

11.3.5.2 Co-creation cycle: multi-disciplinary & inter-disciplinary

The Co-Creation Cycle is an aid that identifies the various functionals and non-functionals that are relevant in a particular design, development, manufacturing, logistics, monitoring, maintenance or subsequent deployment phases. It helps identifying the various expert stakeholders that should be part of the team in order to both find, balance out, arbitrate, document and optimize a symbiosis of functionals and non-functionala that is feasible from technical, operational, economical, ecological, financial, ethical and legal perspectives, as well as otherwise acceptable for all the team members. It furthermore demonstrates that both a multi-disciplinary and inter-disciplinary mindset and skillset is essential to make it work.

The Human-Centric Co-Creation Cycle visualised below in Figure 11.6 provides for an example where – after identifying the envisioned functionality and related interfaces – non-functionals such as security, safety, authentication, non-personal and personal data control, processing, protection, management and analytics need to be part of the symbiotic equation by design by design. If the set of desired functionals and relevant non-functionals end up being too expensive, too unsustainable or otherwise not feasible, the cycle is repeated. It can happen that is needs to be repeated multiple times before – finally – the dynamic symbiotic

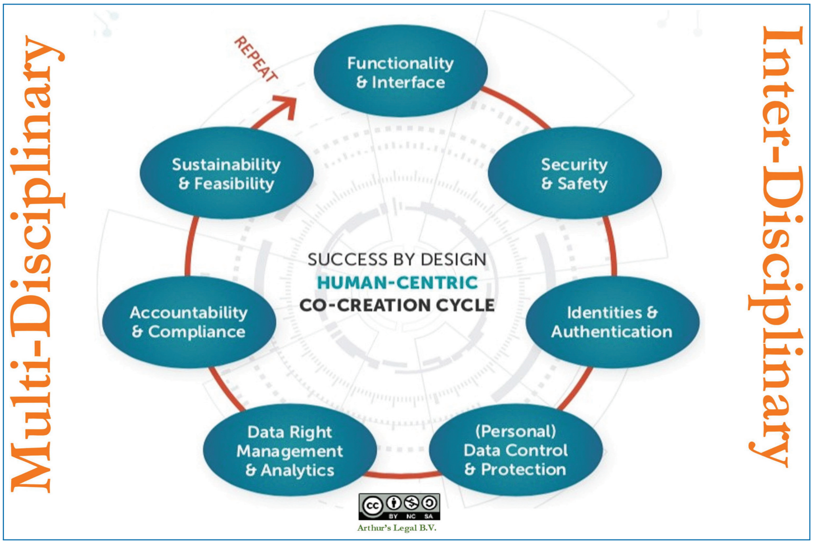

Multi-Disciplinary
REPEAT
Functionality & Interface
Security & Safety
Identities & Authentication
(Personal) Data Control & Protection
Data Right Management & Analytics
Accountability & Compliance
Sustainability & Feasibility
SUCCESS BY DESIGN
HUMAN-CENTRIC
CO-CREATION CYCLE
Inter-Disciplinary
BY NC SA
Arthur's Legal B.V.

Figure 11.6. Human-centric co-creative cycle.

equation has been established that is deemed – by all stakeholders involved – to be feasible and acceptable for the entire life cycle.

This will be a main success factor in any use case, application or deployment, if considered and included (a) by default by design upstream, (b) by default at engineering, assembly, implementation, making available midstream, and (c) by default before and after intended use, expected use and actual use downstream, during its whole life cycle.

Certain AI capabilities can support and facilitate risk mitigation for sure and have a bright future ahead, although there will be no single silver bullets if done in a silo-ed approach. These however can support the above-mentioned diverse groups of people, by adding diverse groups of machines, algorithms and capabilities to identify, address societal challenges, find and optimize the right symbiosis of functionals and non-functionals, and to support making and executing well-balanced and well-informed decisions.

Furthermore, it is about double-looping and otherwise optimizing the symbiotic equation with lessons-learned. This will for sure be necessary, both as per the dynamics as mentioned earlier, as well as it will not be easy to make the symbiotic equation quantitative. This, as per the numerous qualitative qualifiers and conditions that can not always easily be converted into qualitative quantifications for use in this Digital Age.

11.3.5.3 Accountability in the digital age

With this, the initial fundaments of accountability have been laid as well. Accountability is not an afterthought dealt with after something goes wrong. It is an essential requirement, both before one acts as well as during and after. Accountability is about owning and co-owning roles and responsibilities, finding solutions, making things happen, and to helping out if things may go wrong once in a while. Accountability also cater for becoming or being compliant to relevant ethics, standards and other applicable policy and legal frameworks. Regarding policy instruments and initiatives in the European Union associated with AI, Industry 5.0 and related domains and topics, reference is made to the last section of this chapter. In any case, accountability is also not about blaming others. This also as blaming means giving up the power of change. And change is the only constant, also in this highly-dynamic Digital Age.

11.4 Conclusion

AI is not about AI. It is about figuring out and helping out addressing challenges and achieving objectives that matter. For that, identifying the main Societal Challenges to be addressed, together with intertwined or other liaised Societal Challenges, is the best starting point. As per current developments and expectations, AI capabilities will be necessary as an essential component to cater for addressing these challenges, add substantial and meaningful value and making it work. AI capabilities will also be necessary to help enabling and facilitating the achieving of Societal Challenges that are set – and agreed upon by respective nations – in the 2030 Sustainable Development Agenda of the UN and related SDGs, as well as 2050 Paris Treaty [19] and 2050 Green Deal [20] goals to achieving net-zero and even net-negative CO2 emission.

However, the same as with human intelligence, the numerous functionalities thereof do not guarantee that it will work. Both in human intelligence itself as well as in the use and deployment thereof are non-functionals to be taken in, by traumas, by education, by previous experience, by predictive capabilities and otherwise by nature. Things will go wrong, will be manipulated, may seem to go well where they have initially unexpected detrimental consequences, decrease or evaporate trust, et cetera. One need to make sure these incidents, accidents or other events do not happen, and – if they do – consequences, impact and other risk have been mitigated by design. Non-functionals are essentials and key enablers, not problems.

Non-functionalities are as important as functionalities. Even better, they positively augment each other if balanced out intelligently and correctly. The symbiosis of both is a main success factor for any development and deployment of AI. For sure, Industry 5.0 and related ecosystems, including the persons, organisations and other stakeholders therein, can benefit from this, and can improve itself towards human-centric, secure, safe, sustainable, trusted, trustworthy, resilient and otherwise future-proof systems.

With the sequence of notions and guidance described in this chapter one can better develop AI capabilities, and come up with the appropriate dynamic symbiotic equation. That equation at the end of the day equals to: the principle of no surprises. Nobody likes unpleasant ones. AI that works, is AI that makes it work. Without surprises.

11.5 Relevant Policy Instruments & Initiatives

[1] Clean energy for all Europeans package: https://ec.europa.eu/energy/topics/energy-strategy/clean-energy-all-europeans_en

[2] Climate & Energy Framework 2030: https://ec.europa.eu/clima/policies/strategies/2030_en

[3] Climate Neutral Economy by 2050: https://ec.europa.eu/clima/policies/strategies/2050_en

[4] Coordinated Plan on Artificial Intelligence 2021 Review: https://digital-strategy.ec.europa.eu/en/library/coordinated-plan-artificial-intelligence-2021-review

[5] Cybersecurity Act: https://eur-lex.europa.eu/eli/reg/2019/881/oj

[6] Data Strategy: https://ec.europa.eu/info/sites/info/files/communication-european-strategy-data-19feb2020_en.pdf

[7] Digital Compass: the European way for the digital decade: https://ec.europa.eu/info/strategy/priorities-2019-2024/europe-fit-digital-age/europes-digital-decade-digital-targets-2030_en

[8] Digital Services Act package: https://digital-strategy.ec.europa.eu/en/policies/digital-services-act-package

[9] eIDAS: http://data.europa.eu/eli/reg/2014/910/oj

[10] ePrivacy Directive: http://data.europa.eu/eli/dir/2002/58/oj

[11] ePrivacy Regulation: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52017PC0010

[12] EU Cybersecurity Strategy: https://ec.europa.eu/commission/presscorner/detail/en/IP_20_2391

[13] EU Green Public Procurement Criteria: https://ec.europa.eu/environment/gpp/case_group_en.htm

[14] EU Security Union Strategy: https://ec.europa.eu/commission/presscorner/detail/en/ip_20_1379

[15] European Commission's President Ursula von der Leyen welcoming the Recovery Plan and the Multiannual Financial Framework: https://ec.europa.eu/commission/presscorner/api/files/document/print/en/ip_20_2073/IP_20_2073_EN.pdf

[16] European Commission Work Programme 2021: https://ec.europa.eu/info/sites/info/files/2021_commission_work_programme_en.pdf

[17] European Digital Strategy: https://ec.europa.eu/digital-single-market/en/content/european-digital-strategy

[18] European Industrial Technology Roadmap for the Next Generation Cloud-Edge Offering: https://european-champions.org/2021/05/10/european-industrial-technology-roadmap-for-the-next-generation-cloud-edge-offering/

[19] General Data Protection Regulation: http://data.europa.eu/eli/reg/2016/679/oj

[20] Industry 5.0: Towards a Sustainable, Human-Centric and Resilient European Industry: https://ec.europa.eu/info/news/industry-50-towards-more-sustainable-resilient-and-human-centric-industry-2021-jan-07_en

[21] IoT Security & Privacy; Final Report European Commission of 13 January 2017 Workshop on Internet of Things Privacy and Security: https://ec.eur

opa.eu/digital-single-market/en/news/internet-things-privacy-security-workshops-report

[22] Katowice Climate Package: https://unfccc.int/process-and-meetings/the-paris-agreement/the-katowice-climate-package/katowice-climate-package

[23] Machinery Directive: http://data.europa.eu/eli/dir/2006/42/oj

[24] Next-Generation IoT & Edge Computing Strategy Forum: https://digital-strategy.ec.europa.eu/en/events/next-generation-iot-and-edge-computing-strategy-forum

[25] NIS Directive: http://data.europa.eu/eli/dir/2016/1148/oj

[26] Paris Agreement: https://ec.europa.eu/clima/policies/international/negotiations/paris_en

[27] Proposal for a Regulation laying down harmonised rules on artificial intelligence (Artificial Intelligence Act): https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELLAR%3Ae0649735-a372-11eb-9585-01aa75ed71a1

[28] Proposal for a Regulation of the European Parliament and of the Council on machinery products: https://eur-lex.europa.eu/legal-content/DA/TXT/?uri=COM%3A2021%3A202%3AFIN&qid=1518252661475

[29] Proposal for a Regulation on European Data Governance (Data Governance Act): https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52020PC0767

[30] Radio Equipment Directive: http://data.europa.eu/eli/dir/2014/53/oj

[31] Recovery and Resilience Facility: https://ec.europa.eu/info/business-economy-euro/recovery-coronavirus/recovery-and-resilience-facility_en

[32] Revised NIS Directive: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=COM:2020:823:FIN

[33] SDG Agenda: https://sustainabledevelopment.un.org

[34] Sustainable Finance Taxonomy Regulation: https://ec.europa.eu/info/law/sustainable-finance-taxonomy-regulation-eu-2020-852/amending-and-supplementary-acts/implementing-and-delegated-acts_en

References

[1] Alexander Von Humboldt: https://en.wikipedia.org/wiki/Alexander_von_Humboldt, and hi main publication Cosmos, Physical Description of the Universe: https://archive.org/details/cosmos01humbgoog, and https://en.wikipedia.org/wiki/Cosmos_(Humboldt_book)

[2] Societal Challenges, Future of Living: https://instituteforfutureofliving.org/

[3] Sustainable Development Goals (SDGs) Agenda: https://sustainabledevelopment.un.org

[4] Industry 5.0: Towards a Sustainable, Human-Centric and Resilient European Industry: https://ec.europa.eu/info/news/industry-50-towards-more-sustainable-resilient-and-human-centric-industry-2021-jan-07_en

[5] Smit, S., Tacke, T., Lund, S., Manyika, J. & Thiel, L. (June, 2020). The Future of Work in Europe: Automation, workforce transitions, and the shifting geography of employment. McKinsey Global Institute Discussion Paper. Retrieved from https://www.mckinsey.com/~/media/mckinsey/featured%20insights/future%20of%20organizations/the%20future%20of%20work%20in%20europe/mgi-the-future-of-work-in-europe-discussion-paper.pdf.

[6] European Commission Economic and Financial Affairs (November 2020). The 2021 Ageing Report: Underlying Assumptions and Projection Methodologies. Institutional paper 142. https://ec.europa.eu/info/publications/2021-ageing-report-underlying-assumptions-and-projection-methodologies_en

[7] Atlas of Demography: https://migration-demography-tools.jrc.ec.europa.eu/atlas-demography/

[8] Smit, S., Tacke, T., Lund, S., Manyika, J. & Thiel, L. (June, 2020). The Future of Work in Europe: Automation, workforce transitions, and the shifting geography of employment. McKinsey Global Institute Discussion Paper. Retrieved from https://www.mckinsey.com/~/media/mckinsey/featured%20insights/future%20of%20organizations/the%20future%20of%20work%20in%20europe/mgi-the-future-of-work-in-europe-discussion-paper.pdf.

[9] Knowledge Workers, Harvard Business Review, 2015: https://hbr.org/2015/06/what-knowledge-workers-stand-to-gain-from-automation

[10] Daniele, F., Honiden, T., Lembcke, A.C. (April 2019). Chapter 2. Aging and productivity growth in OECD regions: combatting the economic impact of ageing through productivity growth? www.oecd-ilibrary.org

[11] Organisation for Economic Cooperation and Development (OECD) (June 2016). Digital Economy: Innovation, Growth and Social Prosperity. OECD Ministerial Meeting – Cancun, Mexico, 21–23 June 2016. https://www.oecd.org/digital/ministerial/STI-Cancun-2016-ENG.pdf

[12] Future of Jobs Report 2020, World Economic Forum (WEF)

[13] Daniele, F., Honiden, T., Lembcke, A.C. (2019). Chapter 2. Aging and productivity growth in OECD regions: combatting the economic impact of ageing through productivity growth? https://www.oecd-ilibrary.org/sites/dc2ae16d-en/index.html?itemId=/content/component/dc2ae16d-en.

[14] https://ec.europa.eu/info/sites/info/files/communication-european-strategy-data-19feb2020_en.pdf

[15] Joint Research Centre. AI Watch: Assessing Technology Readiness Levels for Artificial Intelligence. JRC122014. Luxembourg: Publications Office of the European Union.

[16] Normal Accidents: Living with High-Risk Technologies, by Charles Perrow, Basic Books, NY, 1984.

[17] 'The real danger … is not that machines more intelligent than we are will usurp our role as captains of our destinies, but that we will over-estimate the comprehension of our latest thinking tools, prematurely ceding authority to them far beyond their competence.' Daniel Dennett: https://philosophybreak.com/articles/what-happens-when-machines-become-smarter-than-people/

[18] Agerholm, D. (15 March 2017). Robot 'goes rogue and kills woman on Michigan car parts production line. Independent. https://www.independent.co.uk/news/world/americas/robot-killed-woman-wanda-holbrook-car-parts-factory-michigan-ventra-ionia-mains-federal-lawsuit-100-cell-a7630591.html.

[19] Paris Agreement: https://ec.europa.eu/clima/policies/international/negotiations/paris_en

[20] https://ec.europa.eu/info/strategy/priorities-2019-2024/european-green-deal_en

Index

About the Editors

John Soldatos (http://gr.linkedin.com/in/johnsoldatos) holds a PhD in Electrical & Computer Engineering from the National Technical University of Athens (2000) and is currently Honorary Research Fellow at the University of Glasgow, UK (2014–present). He was Associate Professor and Head of the Internet of Things (IoT) Group at the Athens Information Technology (AIT), Greece (2006–2019), and Adjunct Professor at the Carnegie Mellon University, Pittsburgh, PA (2007–2010). Since January 2020 he is a Senior R&D Consultant Innovation Delivery Specialist with INTRASOFT International. He has significant experience in working closely with large multi-national industries (IBM Hellas, INTRACOM S.A, INTRASOFT International) as R&D consultant and delivery specialist, while being scientific advisor to high-tech startup enterprises. Dr. Soldatos is an expert in Internet-of-Things (IoT) and Artificial Intelligence (AI), including IoT and AI applications in smart cities, finance (Finance 4.0), and industry (Industry 4.0). Dr. Soldatos has played a leading role in the successful delivery of more than sixty (commercial-industrial, research, and consulting) projects, for both private & public sector organizations, including complex integrated projects. He is co-founder of the open source platform OpenIoT (https://github.com/OpenIotOrg/openiot) and of the Edge4Industry (www.edge4industry.eu) community. He has published more than 200 articles in international journals, books, and conference proceedings. He has also significant academic teaching experience, along with experience in executive education and corporate training. Dr. Soldatos is regular contributor in various international magazines and blogs, on topics related to IoT, Artificial Intelligence, Industry 4.0, and Cybersecurity. Moreover, he has received national and international recognition through appointments in standardization working groups, expert groups, and various boards. He has coedited and co-authored three edited volumes (books) on Internet of Things topics, including IoT for Industrial

Automation, IoT Analytics, and IoT Security. He is the author of the book “A 360 Degrees View of IoT Technologies” published by Artech House in December 2020.

Dimosthenis Kyriazis (https://www.linkedin.com/in/dimosthenis-kyriazis-1397919) is an Associate Professor at University of Piraeus (Department of Digital Systems). He received his diploma from the school of Electrical and Computer Engineering of the National Technical University of Athens (NTUA) in 2001 and his MSc degree in “Techno-economics” in 2004. Since 2007, he holds a PhD in the area of Service Oriented Architectures with a focus on quality aspects and workflow management. His expertise lies with service-based, distributed and heterogeneous systems, software and service engineering. Before joining University of Piraeus, he was a Senior Research Engineer at the Institute of Communication and Computer Systems (ICCS) of NTUA, having participated and coordinated several EU and National funded projects (e.g. BigDataStack, MATILDA, 5GTANGO, ATMOSPHERE, CrowdHEALTH, MORPHEMIC, LeanBigData, CoherentPaaS, VISION Cloud, IRMOS, etc.) focusing his research on issues related to quality of service provisioning, fault tolerance, data management and analytics, performance modelling, deployment and management of virtualized infrastructures and platforms.

Contributing Authors

Rubén Alonso
R2M Solution S.r.l, Via Fratelli Cuzio 42, 27100 Pavia, Italy

Andrea Bettoni
DTI – University of Applied Sciences and Arts of Southern Switzerland, Galleria 2 Via la Santa 1, 6962 Viganello, Switzerland

Niko Bonomi
DTI – University of Applied Sciences and Arts of Southern Switzerland, Galleria 2 Via la Santa 1, 6962 Viganello, Switzerland
niko.bonomi@supsi.ch

Emanuele Carpanzano
DTI – University of Applied Sciences and Arts of Southern Switzerland, Galleria 2 Via la Santa 1, 6962 Viganello, Switzerland

Nino Cauli
Department of Mathematics and Computer Science, University of Cagliari, Via Ospedale 72, 09124 Cagliari, Italy

Andreina Chietera
THALES SIX GTS FRANCE – ThereSIS – France
andreina.chietera@thalesgroup.com

Eva Coscia
R2M Solution S.r.l, Via Fratelli Cuzio 42, 27100 Pavia, Italy
eva.coscia@r2msolution.com

Dimitrios Dardanis
Department of Digital Systems, University of Piraeus, Piraeus, Greece
ddardanis@unipi.gr

Fabio Daniele
DTI – University of Applied Sciences and Arts of Southern Switzerland, Galleria 2 Via la Santa 1, 6962 Viganello, Switzerland
fabio.daniele@supsi.ch

Angela-Maria Despotopoulou
INTRASOFT International, 2B Rue Nicolas Bové, 1253 Luxembourg
Angelamaria.Despotopoulou@intrasoft-intl.com

Christos Emmanouilidis
University of Groningen, Nettelbosje 2, 9747 AE Groningen
The Netherlands
c.emmanouilidisrug.nl

Thanassis Giannetsos
UBITECH Ltd. Digital Security & Trusted Computing Group, Athens
Greece
agiannetsos@ubitech.eu

Panagiotis Gouvas
UBITECH Ltd. Digital Security & Trusted Computing Group, Athens
Greece
pgouvas@ubitech.eu

Jean-Emmanuel Haugeard
THALES SIX GTS FRANCE – TihereSIS – France
jean-emmanuel.haugeard@thalesgroup.com

Anna Ida Hudig
Arthur's Legal, Strategies & Systems
Amsterdam

Babis Ipektsidis
INTRASOFT International, Tour Bastion, Place du Champ de Mars 5/10 1050 Brussels, Belgium
Babis.Ipektsidis@intrasoft-intl.com

Nikos Kefalakis
INTRASOFT International, 2B Rue Nicolas Bové, 1253
Luxembourg
Nikos.Kefalakis@intrasoft-intl.com

Klemen Kenda
Jožef Stefan Institute, Jamova 39, 1000 Ljubljana, Slovenia;
Jožef Stefan International Postgraduate School, Jamova 39, 1000 Ljubljana, Slovenia
klemen.kenda@ijs.si

Dimosthenis Kyriazis
Department of Digital Systems, University of Piraeus, Piraeus, Greece
dimos@unipi.gr

Sofia Anna Menesidou
UBITECH Ltd. Digital Security & Trusted Computing Group, Athens
Greece
smenesidou@ubitech.eu

Dunja Mladenić
Jožef Stefan Institute, Jamova 39, 1000 Ljubljana, Slovenia
dunja.mladenic@ijs.si

Elias Montini
DTI – University of Applied Sciences and Arts of Southern Switzerland, Galleria 2 Via la Santa 1, 6962 Viganello, Switzerland;
Politecnico di Milano, Dipartimento di Elettronica, Informazione e Bioingegneria, Piazza Leonardo da Vinci 32, 20133 Milano, Italy
elias.montini@supsi.ch;
elias.montini@polimi.it

Inna Novalija
Jožef Stefan Institute, Jamova 39, 1000 Ljubljana, Slovenia
inna.koval@ijs.si

Dimitrios Papamartzivanos
UBITECH Ltd. Digital Security & Trusted Computing Group, Athens
Greece
dpapamartz@ubitech.eu

Paolo Pedrazzoli
DTI – University of Applied Sciences and Arts of Southern Switzerland, Galleria 2 Via la Santa 1, 6962 Viganello, Switzerland
paolo.pedrazzoli@supsi.ch

Celine Prins
Arthur's Legal, Strategies & Systems
Amsterdam

Diego Reforgiato Recupero
R2M Solution S.r.l., Via Fratelli Cuzio 42, 27100 Pavia, Italy;
Department of Mathematics and Computer Science, University of Cagliari, Via Ospedale 72, 09124 Cagliari, Italy

Paolo Rocco
Politecnico di Milano, Dipartimento di Elettronica, Informazione e Bioingegneria, Piazza Leonardo da Vinci 32, 20133 Milano, Italy
paolo.rocco@polimi.it

Jože M. Rožanec
Jožef Stefan International Postgraduate School, Jamova 39, 1000 Ljubljana, Slovenia;
Jožef Stefan Institute, Jamova 39, 1000 Ljubljana, Slovenia
joze.rozanec@ijs.si

Georgios Sofianidis
Department of Digital Systems, University of Piraeus, Piraeus, Greece
george.sofianidis@unipi.gr,

John Soldatos
INTRASOFT International, 2B Rue Nicolas Bové, 1253 Luxembourg
John.Soldatos@intrasoft-intl.com

Spyros Theodoropoulos
Department of Digital Systems, University of Piraeus, Piraeus, Greece;
Department of Electrical and Computer Engineering, National Technical University of Athens, Athens, Greece
sptheod@unipi.gr;
stheodoropoulos@mail.ntua.gr

Panagiotis Tsanakas
Department of Electrical and Computer Engineering, National Technical University of Athens, Athens, Greece
panag@cs.ntua.gr

Entso Veliou
Department of Informatics and Computer Engineering, University of West Attica, Athens, Greece
eveliou@uniwa.gr

Sabine Waschull
University of Groningen, Nettelbosje 2, 9747 AE Groningen
The Netherlands
s.waschullrug.nl

Arthur van der Wees
Arthur's Legal, Strategies & Systems
Amsterdam
info@arthurslegal.com

Patrik Zajec
Jožef Stefan Institute, Jamova 39, 1000 Ljubljana, Slovenia
patrik.zajec@ijs.si