9

Ethical Considerations and Trustworthy Industrial AI Systems

Ovidiu Vermesan[1], Cristina De Luca[2], Reiner John[3], Marcello Coppola[4], Björn Debaillie[5], and Giulio Urlini[6]

[1]SINTEF AS, Norway
[2]Silicon Austria Labs GmbH, Austria
[3]AVL List GmbH, Austria
[4]STMicroelectronics, France
[5]IMEC, Belgium
[6]STMicroelectronics, Italy

Abstract

The ethics of AI in industrial environments is a new field within applied ethics, with notable dynamics but no well-established issues and no standard overviews. It poses many more challenges than similar consumer and general business applications, and the digital transformation of industrial sectors has brought into the ethical picture even more considerations to address. This relates to integrating AI and autonomous learning machines based on neural networks, genetic algorithms, and agent architectures into manufacturing processes.

This article presents the ethical challenges in industrial environments and the implications of developing, implementing, and deploying AI technologies and applications in industrial sectors in terms of complexity, energy demands, and environmental and climate changes.

It also gives an overview of the ethical considerations concerning digitising industry and ways of addressing them, such as potential impacts of AI on economic growth and productivity, workforce, digital divide, alignment with trustworthiness, transparency, and fairness.

Additionally, potential issues concerning the concentration of AI technology within only a few companies, human-machine relationships, and behavioural and operational misconduct involving AI are examined.

Manufacturers, designers, owners, and operators of AI—as part of autonomy and autonomous industrial systems—can be held responsible if harm is caused. Therefore, the need for accountability is also addressed, particularly related to industrial applications with non-functional requirements such as safety, security, reliability, and maintainability supporting the means of AI-based technologies and applications to be auditable via an assessment either internally or by a third party. This requires new standards and certification schemes that allow AI systems to be assessed objectively for compliance and results to be repeatable and reproducible.

This article is based on work, findings, and many discussions within the context of the AI4DI project.

Keywords: Artificial intelligence, ethics, digitising industry, industry- grade AI, industrial internet of things, machine ethics, explainable AI, trustworthiness, responsible AI, technology ethics.

9.1 Introduction

We all remember the frequent quotation of Isaac Asimov and his famous Three Laws of Robotics (1942). First Law: A robot may not injure a human being or, through inaction, allow a human being to be harmed. Second Law: A robot must obey orders given by humans, except where such orders conflicts with the First Law. Third Law: A robot must protest its own existence, provided such protection does not conflicts with First and Second Law. They perfectly reflects the need for future use of AI not to harm human beings. On the other side, the definition of AI proposed in the European Commission's Communication on AI [2][3][4][5] states that "Artificial intelligence (AI) refers to systems that display intelligent behaviour by analysing their environment and taking actions – with some degree of autonomy – to achieve specific goals. AI- based systems can be purely software-based, acting in the virtual world (e.g., voice assistants, image analysis software, search engines, speech and face recognition systems), or AI can be embedded in hardware devices (e.g., advanced robots, autonomous vehicles, drones or Internet of Things (IoT) applications)."

In the context of the AI4DI project under European Union's Horizon 2020 research and innovation programme [1], AI is defined as a machine's ability

to perform logical analysis, acquire knowledge, and adapt to an industrial environment that varies over time or in context. These abilities include the collective attributes of a machine (i.e., computer, robot, or intelligent IoT device) to perform functions such as perception, understanding, reason, prediction, learning, decision making and action.

Another definition [6] mentions that AI is an activity dedicated to creating machine intelligence. Intelligence is a quality that allows an entity to function appropriately and with insight and foresight in its environment.

The increased number of intelligent machines, products and services, (i.e., equipment, industrial IoT devices with embedded AI, etc.), based on machine learning (ML), artificial neural networks (ANNs) and deep learning (DP), deployed in industrial environments, require to open the discussion on ethical principles and how these relate to AI.

AI is defined based on outcomes and actions [23]. The ethics of AI in industrial environments are evolving due to discussions around industrial AI trust, technical problems that focus on achieving the desired outcome for AI-based technologies and applications in manufacturing sectors. This is a new field within applied ethics and comes with notable dynamics, controversial issues, a lack of standards and no common agreement on principles about ethics.

Trust in an industrial AI system has multiple dimensions combining system dependability characteristics (e.g., privacy, security, safety, reliability, availability, resilience, connectability and maintainability) with human and machine behaviour. There is a need for a greater understanding of how individuals interact with machines and how machines/things interact with other machines/things to extend trust.

Trust in industrial AI systems is a characteristic of human-to-machine and machine-to-machine relationships formed with different industrial AI-based systems. In industrial processes, a further understanding of how individuals interact with AI-based machines and how these machines/things interact with other machines/things is critical for building the industrial AI trust concept. In many industrial processes, AI trust is developed by considering the performance (e.g., accuracy, robustness, stability, speed, data quality, etc.) of AI, the ML model, the operations (compliance, dependability, response to uncertainty, monitoring, governance, etc.) of the industrial AI system and the set of rules, guidelines, and standards (e.g., ethical, technical, etc.) in the industrial workflow. The rules/guidelines related to, for example, transparency, explainability, bias, and fairness apply to both the design of the industrial AI system, how it is used and how its functions are explained in the industrial process.

Uncertainty and vulnerability are two of the core elements of AI trust. In addressing industrial AI trust issues, industrial stakeholders must select strategies that reduce uncertainty or decrease vulnerability, depending on the context of the problems. Design for industrial AI trust requires evaluating the operating assumptions and examining how those assumptions can function to put some users of the AI system at risk. Understanding and designing AI trust systems require an understanding of the rules of the AI system and the functions of autonomous/cognitive elements.

From an industrial AI technology perspective, trust refers to trust measurement capabilities. This requires the use of trust assessment approaches, such as recommendation and reputation systems, which calculate the trustworthiness of one industrial AI system to match it against the need for trust of another industrial AI system.

As industrial AI technologies are maturing and AI-based applications are proliferating in different industrial sectors, new standards are demanded that describe measurable and testable levels of transparency are required; in this way the AI-based systems are objectively assessed for compliance to be reliable, safe, trustworthy, and operate with integrity.

The developed economies understand the game-changing nature of AI and have embraced different approaches to accelerate and control the development of AI technologies and applications.

Industrial AI depends on addressing the trade-off between incorporating the benefits and mitigating the potential disadvantages of AI by simultaneously avoiding the misuse and underuse of AI technologies in industrial environments.

Embracing an ethical approach to industrial AI provides what is considered a twin benefit of using ethics to allow industrial organisations to take advantage of AI's value and anticipate, avoid, and minimise expensive missteps and errors.

A framework of industrial AI principles is based on statements of the values or principles that guide the development and deployment of AI in society and that have already been proposed by different multi-stakeholder organisations and initiatives [22][23][24][25][26][27][28][29]. Asilomar principles [26] provide the greatest number of such principles organised under three issues: research, ethics and values, and longer-term issues. Regarding these principles five topics emerge as key for AI ethics:

- Autonomy as the element to use for whether or not to delegate.
- Beneficence as related to doing only good and providing a benefit.
- Non-maleficence as related to causing no harm and damages.

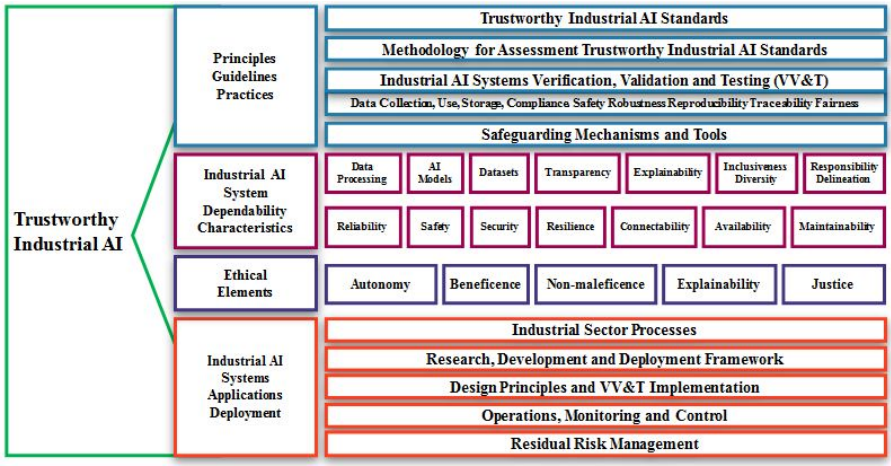

Figure 9.1 A framework for trustworthy industrial AI systems.

- Explainability as related to how the AI-based system does its work and who is responsible for the way it works.
- Justice as related to promoting fairness, and prosperity, preserving solidarity and avoiding unfairness, bias, and discrimination.

A framework for trustworthy industrial AI systems including the elements and principles presented above is illustrated in Figure 9.1.

One key opportunity for the industrial sector in Europe to be competitive is to ensure the take-up of AI technology across its industry. The development of higher efficient electronic components and systems, circuits specifically built to run AI operations (neuromorphic circuits), high-performance computers, quantum technologies and technologies for mapping the human brain accelerate the possible applications of AI-based technologies in industrial sectors and urgently require addressing the issues of ethical challenges that the AI brings.

9.2 Ethics and Responsible AI in Industrial Environments

Nowadays, AI is impacting many aspects of industrial activities. There is a need to understand how AI should be designed to i) operate responsibly, ii) meet stakeholders' expectations and iii) applicable regulations and concerns relating to reliability, privacy data leakages, information transparency, explainability and ethical considerations [16].

Addressing these ethical dilemmas and concerns when developing industrial AI-based solutions strengthens a manufacturer's credibility for delivering products and services and enhances an organisation's reputation in the marketplace. Nevertheless, this is not an easy task, as industrial applications have much higher requirements (e.g., reliability, verifiability, safety, etc). than AI-based products designed for the consumer market.

Many industrial companies address the ethical and environmental concerns around the responsible use of AI in corporate social responsibility strategy to make socially significant decisions and consider using "ethical algorithms" to reduce the risk of unethical behaviours.

There is no single definition for what responsible AI means, and organisations will usually develop their terminology and methodology. Nevertheless, designing AI to operate responsibly means at its core following design principles that allow AI systems to justify and be held responsible for their decisions. In industrial environments, this ultimately comes down to allowing human inspection of the functionality of AI algorithms and models. The development of AI systems is complex, involving many sub-systems with different ethical considerations, making it challenging to inspect and evaluate such systems. The complexity arises from the fact that ethically compliant sub-systems do not necessarily make the overall system ethically compliant. The subsystems interact with each other and exchange feedback, which may change conditions in the application's environment, conditions that cannot always be anticipated during development. This may be the case with AI systems that continue to learn after deployment. Therefore, re-evaluation of ethical compliance must be conducted regularly or with every change of the application context, especially in AI systems with widespread or profound ethical issues. Safety-critical systems, where industry regulation would make the re-evaluation mandatory, can be such a case. A schematic representation of the elements in the development process of AI systems is illustrated in Figure 9.2.

In this context, it is essential to note that designing and developing responsible AI is not a one-time process but rather entails continuous striving to maintain responsible AI systems and keep up with technological advances that may bring new ethical implications.

9.3 Requirements for Industry-Grade AI

Defining the requirements for industry-grade AI is crucial as advanced machines and Industrial Internet of Things (IIoT) devices with enhanced AI

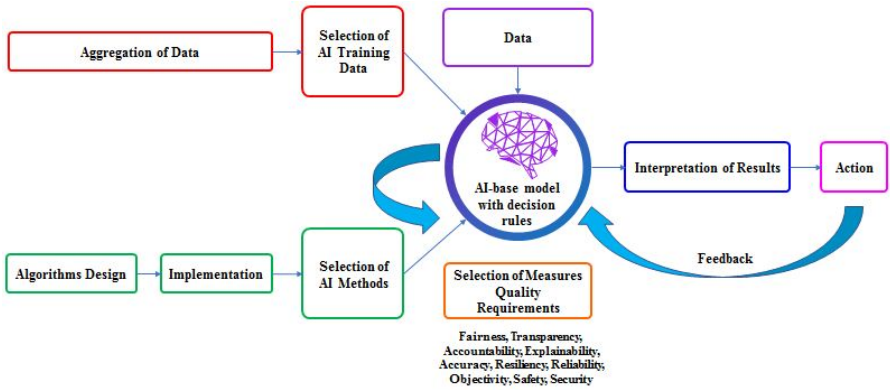

Figure 9.2 Complexity of applicability of ethical considerations resulting from the interaction of subsystems.

capabilities may operate in ways that were not envisaged when the AI-based system was designed and put into operation.

The requirements for industry-grade AI technologies and applications identified by the AI4DI project [1] are illustrated in Figure 9.3. A short

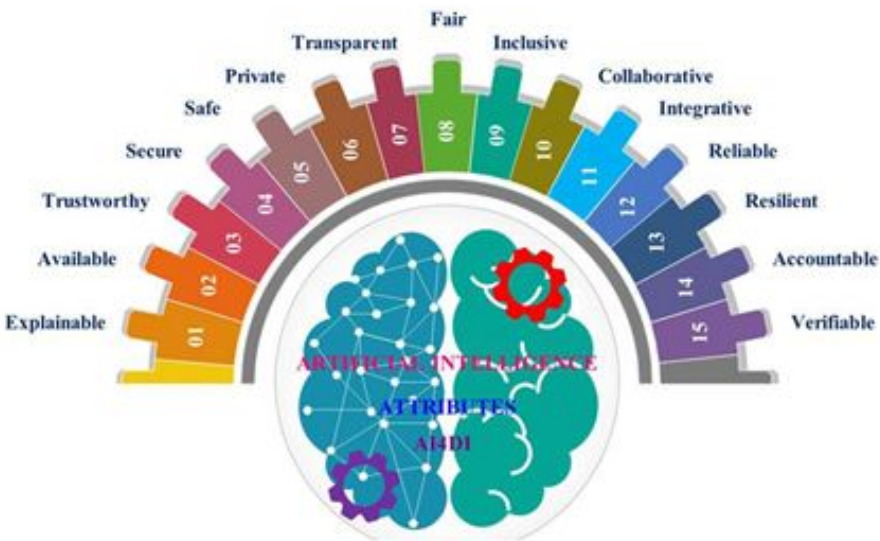

Figure 9.3 Requirements for industry-grade AI [1].

explanation of these requirements is presented in the next paragraphs. Some of these requirements are further addressed in specific sub-sections.

Explainable: Humans must comprehend the decision of AI systems to track down failure and assess decisions, and the AI systems must either provide enough information required to explain its actions and decisions or possibly even explain the output itself (explainable AI).

Available: Industrial applications for AI will target mission-critical tasks along the complete production line, and system outages will have a direct economic impact. Industry-grade AI systems, therefore, need to fulfil high availability standards. In a second step, they should also perform autonomously via online learning over their lifetime to avoid maintenance. Moreover, they should also be quickly available in terms of integration into new applications and process steps.

Trustworthy: When more and more AI-enabled devices become connected through the IIoT, trustworthiness will become an indispensable requirement for AI systems. It is essential that the identity of every AI system can be verified, and that vulnerabilities and inconsistencies are immediately reported.

Secure: Industry-grade AI must implement security measures to ensure robustness against all types of attack vectors through different devices, workers, operators, etc. This also includes securing the AI system by making it robust against adversarial attacks and manipulated input data. The communication between edge computing devices needs to be secured with encryption and authentication mechanisms. Security is very important, particularly when we execute AI on the edge. Therefore, protecting embedded ML models against attacks by safeguarding the integrity (fooling decision) and confidentiality of the data is vital when IIoT systems are deployed in the field. With AI, the overall attack surface is large since we are gathering algorithmic attacks (such as adversarial examples) and physical attacks (side-channel and fault injection analysis). To address these issues, it is important to check the robustness of models, add cryptographic-based authentication schemes and add secure boot-like technologies to enforce the trust of embedded AI systems against malicious tampering.

Safe: AI systems that operate physically next to and collaboratively with humans through robots or other machines must comply with current and future safety standards to prevent accidents. Notably, the employed AI systems must be robust against implausible data and operate with extremely low

latency to quickly react to unforeseen events. Likewise, AI for the control of safety-critical processes must also comply with the latest safety standards.

Private: Industry-grade AI will operate on mission-critical personal data from the manufacturer and customers and business-critical information (data security, confidentiality, etc). This data must be kept confidential and protected from external access. This precludes external cloud storage and the application of typical big data methods. Instead, information must be processed locally at the edge and only leverage data available within privacy limits (smart data).

Transparent: The state, actions, and decisions of an AI system must be inspectable and understandable at any point in time. This will be supported by digital twins that represent the complete system state at any point in time. AI methods for data visualisation can further enhance transparency and make the systems state easier to understand.

Fair: AI technologies that support or automate decision processes must adhere to the same fairness and compliance standards as defined by the industrial sector regulation.

Inclusive: AI systems need to include humans and existing systems in their operations to avoid the formation of isolated non-AI capable sub-systems within a process, production system or supply chain.

Collaborative: Industry-grade AI will not be concentrated on a single device or system. Instead, many different AI-enabled sub-systems will be distributed (distributed AI) across IoT nodes, embedded devices, and other edge devices (AI-Born embedded systems). These devices need to self- organise and collaborate to ensure coherent operation at the level of the whole system. They also need to collaborate with humans physically (e.g., human-robot collaboration) and by exchanging information (human-machine interfaces).

Integrative: Industry-grade AI systems must be open and flexible to ensure that they can be integrated seamlessly into existing systems and processes. This is a key prerequisite of establishing AI methods in the industry according to a sustainable roadmap.

Reliable: Reliability and dependability (dependable AI) are key prerequisites for AI systems that are put into continuous operation with short maintenance time in mission-critical production environments. AI must not harm productivity by an unreliable operation that requires regular human intervention or even causes system outages.

Resilient: Industry-grade AI must remain stable even when other parts of the process fail. In the future, they should even be able to detect the failure and initiate measures for compensating it.

Accountable: Industry-grade AI that supports or even replaces human decisions must be implemented to ensure that it can be made accountable for its output (e.g., via the supplier of the AI system).

Verifiable: AI systems for industrial applications must fulfil the same standards as legacy systems and will be applied to safety-, mission-, and business-critical tasks. This requires that Industry-grade AI systems can be validated (to reach correct results), verified (verifiable AI) and certified (certifiable AI) for the targeted applications.

9.4 Industrial AI Challenges

The industrial AI technologies are powered by complex programming and algorithms run on high-performance energy-consuming computing units. Hence, the AI technologies are affecting how we interact with the environment and the resources used to power the different AI-based solutions.

While AI has many positive impacts, the widespread use of AI solutions in industrial environments can have indirect adverse and hidden effects that can harm the environment.

In many cases, phrases like "the data is the new oil" are used to highlight the digital transformation without considering that as for the oil, when data is used excessively, it is polluting the environment.

Raw data has no value in itself. Instead, the value is created when collected effectively and accurately, connected to other relevant data, done on time, processed, and refined. When well refined, usable data immediately becomes a decision-making tool – information – allowing companies to use it in the manufacturing decision-making and process automation.

Processing the information requires advanced AI-based hardware accelerated devices, software, algorithms, model, storage, computing, and connectivity capabilities that increase the complexity of the systems, i.e. use more natural resources, energy, and pollute the environment and generate new waste.

Large-scale deployment of AI could have both positive and negative impacts on the environment. Positive impacts can improve the user experience and the durability of machines. Thanks to preventive maintenance, better products can be made by adapting the production process to external

situations. Negative impacts include increased complexity, use of natural resources, pollution, waste, and energy consumption.

In the following paragraphs, these challenges are highlighted to present a comprehensive overview of the trade-off that must be considered when developing AI-based IIoT and other types of systems in industrial environments.

9.4.1 Complexity

Ethical implications and challenges are present even in the simplest AI systems. In many cases, the more complex AI systems, the greater the challenges associated with their unpredictability and lack of transparency.

Industrial AI-based solutions result from multidisciplinary cooperation, and almost all AI-based systems are complex systems integrating IIoT devices, hardware, software, models, algorithms, and platforms.

The robustness and performance of models and algorithms are strongly dependent on their learning abilities; hence, improving learning ability performance will increase complexity. For instance, in the case of deep neural networks, widening or deepening the network will enhance the learning ability and the performance of the overall AI-based solution, but in many cases, it will also increase its complexity. The many internal hidden layers will be more challenging to penetrate for the purpose of analysis, including verification of ethical compliance.

In industrial environments, complex problems have multiple layers, each of which has multiscale parameters and characteristics, with the different layers correlated to each other. The AI-based industrial complex systems consist of numerous elements/components with a spatiotemporal multiscale structure between the system and elements/components scale due to the collective effect of these factors.

The complexity of the AI-based systems in industrial environments is in many cases determined by the difference between intelligent machines and human thinking. As demonstrated in many AI applications, statistical methods of ML, including the field of neural networks, vary in many forms from biological concepts of understanding, thinking, decision-making or learning.

9.4.2 Use of Natural Resources

Large-scale deployment of AI could have both positive and negative impacts on the environment. AI is creating positive environmental impacts in many applications but can negatively impact others where extensive use of natural

resources has already damaged the environment. By increasing the demand for natural resources to increase automation and yield, AI can accelerate environmental degradation.

Data used in AI applications must be captured, stored, analysed, and transferred to different locations, which requires significant amounts of processing power. It is estimated that 175 zettabytes (ZB) worth of data will be stored globally by 2025 [18], which means extensive use of natural resources to address the need of energy, cooling, water, buildings etc.

AI will probably increase the demand for new materials needed for batteries that power the devices based on AI algorithms to perform intelligent functions on the manufacturing floor.

9.4.3 Pollution and Waste

AI is being used in applications to combat waste pollution. However, as more companies across more industries begin to use AI, there is growing concern that AI technologies will also extend the climate crisis.

AI and ML algorithms are training for longer and longer, using more and more sensors/devices generating data and consuming more and more energy, thus directly or indirectly increasing pollution by generating more and new types of waste that can be detrimental to the environment.

The use of power-intensive GPUs, energy-inefficient algorithms, a large amount of data to run ML training are all considered contributing to increased carbon dioxide emissions.

In this context, researchers are proposing ways to monitor the carbon footprint of AI algorithms and evaluate the pollution generated by AI applications. Code can be attached to the AI models and algorithms to track the energy use of individual AI-based processing units. An online calculator tool is used by [13] to give the raw carbon emissions produced and the approximate offset carbon emissions (depending on the grid used by the cloud provider).

9.4.4 Energy

AI requires extensive amounts of energy for manufacturing and training/learning. This will increase the carbon footprint of manufacturing products based on AI and the overall energy consumption across the entire lifetime of a product that needs continuously retraining and learning.

Training AI algorithms is an energy-intensive process, and estimates hint that the carbon footprint of training AI is as much as 284 tonnes of carbon

dioxide equivalent, which represent five times the lifetime emissions of an average personal vehicle [12], [14].

In most AI-based solutions, the energy-inefficiency of AI algorithms begins with the need to fine-tune the model for particular tasks, translate from one language to another and perform many iterations until the expected results and performance are obtained.

For AI-based technologies and applications, the solutions to energy consumption issues are using renewables to power the computing capabilities responsible for processing, storing, and training data, distributing the processing and analytics at the edge, designing more energy-efficient algorithms, software/hardware systems, and connectivity (e.g., cellular, wireless).

9.5 Ethical Considerations for Digitising Industry

Digital ethics offers a critical reflection about the changes in the industry and manufacturing processes shaped by digital and AI technologies. The digital divide extends from a technical phenomenon to broader ethical issues related to free competition, economic monopolies, silos that can affect the industrial environments.

Autonomous and intelligent AI-based systems have become more pervasive and designed to reduce human intervention in industrial processes and accelerate automation.

In this context, new ethical considerations must be addressed, and topics such as trustworthiness, fairness, transparency, accountability, explainability, and control must be discussed to develop guidelines, standards, and embedding norms in AI-based systems to support their governance in industrial environments. The following paragraphs offer a short overview of these topics and the challenges linked to AI-based systems.

9.5.1 AI Trustworthiness

Digital technologies in manufacturing are pervasive, and AI trustworthiness is imperative for the manufacturing processes to work correctly. To provide risk-free and reliable operation, intelligent machines and processes require continuous supervising ML algorithms used to make decisions. The control and supervision require essential time and resources, to the point that using digital technologies could become very expensive. On the other hand, not controlling the AI-based process may lead to severe risks for the safety and security of the entire production line. AI trustworthiness is based on technology robustness, bias, fairness, transparency and explainability.

The "EU Ethics guidelines for trustworthy AI" [4] provides high- level requirements/principles for trustworthy systems: Human agency and oversight (empowering human beings in order to make informed decisions; keeping the human in the loop), technical robustness and safety (resilient, reliable system functioning), privacy and data governance, transparency, diversity, non-discrimination and fairness, societal and environmental well-being and lastly, accountability (responsibility and accountability for AI systems).

According to NIST, an AI application's trustworthiness value is derived from several variables such as accuracy, explainability, resiliency, safety, reliability, objectivity, security, and accountability [18].

Several of these variables are addressed in the following sub-sections.

9.5.2 Bias and Fairness

Considering that more decisions are delegated to AI in industrial processes, it is crucial to ensure that the decisions and findings are free from bias and unfairness.

Biases prevent AI applications from making fair decisions in the same way as biases affect humans, and they can reside in both the AI training data and the algorithms, both of which are generated by humans.

Data sets can often contain hidden biases due to being incomplete and not covering the whole ground; in other cases, data sets can originate from sources outside the organisation, exhibiting slightly different ethical values.

Developers may also unintentionally programme biases into AI systems, although this is less often the case in industrial environments than in the consumer market.

In many cases, it is impossible to know in the design phase what algorithms based on neural networks are learning when they are trained with a specific data set. In industrial processes, the selection of the training sets, the test sets, and the verification and validation of the results to assess the efficiency and fairness of different algorithms are part of accepting or rejecting the use of the algorithms in the industrial process.

Fairness requires knowing why an AI-based automated process made a particular decision and the mechanisms that may change the decision and is thus connected to the AI models' interpretability and transparency of the training, design, development, and deployment processes with which the models were created.

The absence of fairness that results from the performance of an AI-based industrial system is in most cases due to algorithmic bias generated by a particular categorical distinction.

In this context, in industrial environments, it is critical to identify the root cause for introducing bias in AI systems, if any, and how it can be prevented throughout the lifecycle of the AI-based solution.

AI bias in industrial environments - whether in AI algorithms or training data - can promote distrust and generate distorted outcomes, which decreases the potential of AI for the industry. Introducing AI-based solutions in industrial sectors ensures that AI technologies strengthen human decision making. The industry's stakeholders aim to support scientific advancement and standards that can minimise AI bias.

9.5.3 Transparency

Implementing trustworthy AI-based solutions in industrial environments is closely related to some of the other elements presented in this section, such as fairness, accountability, and transparency.

Transparency relates to the capability of an AI system to, always, be able to provide a satisfactory explanation for its decisions, auditable either by an in-house or an independent human authority assessment. In the case of failure causing harm, it should be possible to ascertain why.

AI transparency must be addressed over the lifecycle development of an AI-based solution from the concept, design, deployment, operation, maintenance, upgrade/update and disposal. In approaching AI transparency in many cases, algorithmic transparency and algorithmic decision-making are the starting point.

In industrial environments, several AI components can be based on black-box solutions. To achieve AI transparency, the openness of the development process must be considered when designing AI-based solutions to allow for explainability concerning interpretability and trust in the AI-based systems.

9.5.4 Accountability

The assumption that human beings are the ultimate decision-makers is one of the fundamental premises most laws and regulations rely on when attributing responsibility. As AI-based autonomous devices become more advanced and ubiquitous, that will increasingly be less true when the "decision-maker" is a machine and not a person [8].

In industrial processes using AI technologies and applications, the responsibility for the AI's action/inaction/malfunction is attributed to an actor that is part of a business agreement, the owner, designer/developer, manufacturer, operator of an AI technology or application.

As the autonomous systems develop and become more intelligent new decisions can be made by the AI-based system, and the intelligent machines hold a certain level of responsibility for their actions. A responsibility gap is created when the behaviour of AI-based products deviates from the initial programming of the developer/designer to become a product of its interactions with its environment making the ascription of responsibility highly complex and unclear [7].

9.5.5 Explainability

In industrial environments, ethical concerns may arise when inaccurate and even incorrect predictions are reached related to either the product or the process. To address these concerns, industrial AI developers need to be able to explain how algorithms predict using various technical approaches and the factors that impact the decisions.

The AI-based technology used in industrial processes must explain WHAT it was designed to do, HOW it was designed to do such functions, and WHY it was designed in that distinct way instead of some other way.

Ensuring that AI-based hardware, software, and algorithms do what are intended to do and that there are no biases or unintended consequences must be addressed through validation and evaluation of the AI-based solutions during development by measuring the performance of an AI-based system through implementation to detect bugs, biases, and incorrect assumptions.

AI-based industrial systems can miss essential facts about the environment, and it is crucial to verify that these systems are operating as intended., including whether the AI models accurately estimate what they are supposed to.

AI explainability should be formulated for different systems, such as sensing, perception, and decision-making. Assessment of industrial AI explainability and explanations needs to be aligned with the industrial context, benchmarking, and targeted use cases, applications, or stakeholders (e.g., developers, users, consumers, etc.). High-level requirements for AI explainability need to be defined by industrial regulations or international standards. They should be aligned with the definition of transparency and verifiability for AI applications in various industrial contexts and at different cognition levels.

9.5.6 Control

Control is another matter that impacts trust, explicitly concerning how much control to exercise over AI, ranging from complete human control to complete AI autonomy. Balancing these two extremes is always possible, so the question is rather what form of control can be exercised and how it can be exercised without hampering the benefits of AI.

One approach is to build self-assessment capability into the AI system before deployment to enable the system to take corrective actions during operation, if necessary, even shutting itself down if harm is anticipated.

The idea to control AI-based technologies is to make ongoing self-evaluations and to test an integral part of a system's operation to diagnose how the AI-based system performs and correct any errors.

Ethical data sets could be used to continuously monitor and check for deviant behaviour, implementing an effective and observant response to ethical behaviour deviations of the algorithms.

Another approach is to keep humans in the loop able to intervene and override decisions that may cause harm.

9.5.7 Human-Machine Interaction and Manipulation of Behaviour

When developing human-machine relationships on the manufacturing floor, it is challenging to prognosticate the psychological effects of forming relationships with different intelligent machines.

Straightforward collaboration between humans and machines in industrial environments requires the interactions to be intuitive, seamless, and unobtrusive. This must be reflected in the implementation of AI-based interfaces built to control and manage these interactions.

Relationships with machines may affect human users' mental and social development and create barriers for humans in understanding the relationships between machines.

The cooperation of mixed groups of machines and humans in automated production lines can affect the performance of groups and the perception of their efficiency.

As technology advances, AI algorithms used in industrial processes can develop capabilities to manipulate human behaviour - to identify and exploit human practices, weaknesses, and vulnerabilities. Algorithms can detect the feelings of the humans involved in the production, including fear, disgust, joy, and relaxation.

In industrial environments, this concern is not related to AI taking over but instead aims to raise awareness of the risks involved when human decision- making has been tampered with. For example, a risk is present if an AI machine or an IIoT smart device no longer operates efficiently because this critical element in AI-powered machines has not been set as a goal.

AI-based applications in industrial environments must consider a risk-based approach and differentiate AI uses according to whether they create an unacceptable risk, a high risk or a low risk. The risk of manipulating behaviour is unacceptable if it poses a clear threat to the regular operation of the manufacturing process, security, quality of the outcome and personal safety involved.

9.5.8 Autonomous Industrial Systems

Advances in industrial automation systems and AI have brought autonomous systems into the focus of digital ethics as intelligent machines that can adapt to the environment are strongly interconnected with autonomy.

Machine autonomy in industrial processes is related to the absence of human intervention. Autonomy is characterised by the ability of an AI-based system to make decisions and justify its actions based on its sensing capabilities to adapt to changes, which occur within the system itself, other systems it interacts with, its operation environment, or in the given task.

Autonomous industrial systems can perceive their environment via sensing perception capabilities, create a plan of action according to the context or scenario-related constraints and execute the planned actions safely and reliably using intelligent actuators.

The autonomous industrial systems have characteristics related to process execution, adaptability, self-governance, self-contentedness, and the corresponding abilities that can connect with non-functional requirements for the AI-based system.

The non-functional requirements or capabilities of autonomous industrial systems are interlinked with the system's skill to perform different tasks. The abilities needed to give the AI-based system the characteristics of an autonomous industrial system differ from case to case and depend on the context.

Autonomous systems must operate without the intervention or assistance of human operators and within the requirements defined by the industrial ethical framework.

The applicability of ethical considerations needs to consider the various aspects of inherent decision-making autonomy, mitigation in abnormal situations, and communication with other machines and with human operators.

A simplified high-level reference architecture for AI-based autonomous systems in industrial environments is illustrated in Figure 9.4. The simplified high-level reference architecture is used in the ECSEL AI4DI [1], ArchitectECA2030 [30], AI4CSM [31] to provide an overview and organise the AI-based systems and their functions.

Communication with the environment relies on sensors for observing the environment and actuators for changing environmental conditions to achieve the objectives. Communication and collaboration with humans and other machines in the industrial environment provide information and feedback on the performance and actions of the system. In abnormal situations, capabilities for cognitive information processing allows the system to fall back to a safe operating state or to hand over control to a human operator and take it back when the situation is normal again.

These capabilities rely on mechanisms for self-regulation controlling the various modules, including knowledge bases, and are constantly adapted

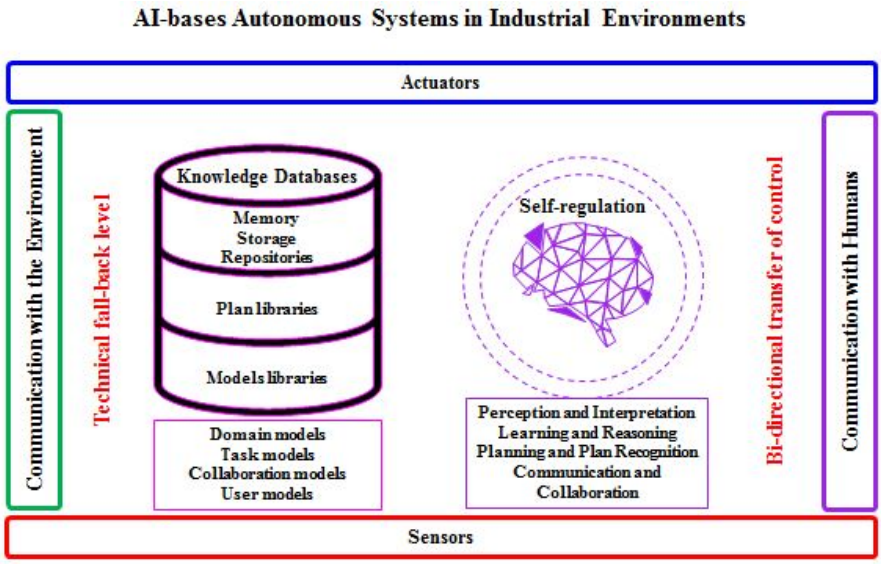

Figure 9.4 Reference architecture for AI-based autonomous systems in industrial environments.

through perception and regulation, learning and reasoning, planning, and plan recognition.

As autonomous industrial systems assume more responsibility in industrial processes in the circumstances previously overseen by human judgment, it is compelling to consider the associated ethical implications. These reflections should include analysis of ethical issues from multiple perspectives, including those of machines and intelligent IIoT devices' designers, operators that control the intelligent machines, and the machines themselves that need to act with ethical correctness.

9.5.9 Machine Ethics

Debates about machine ethics are not new, nor are the arguments about whether AI machines have obligations and rights as humans and animals do. The topic has revived in recent years; questions surrounding AI accountability and autonomy have begun to be addressed more rigorously. The question of whether AI is responsible for its own decisions, actions and consequences, or whether this responsibility falls to the humans that design, develop, operate and assess AI can no longer be answered directly. Many interpretations and nuances are involved in answering this question.

Machine ethics includes how humans design, build, use, and treat AI-based machines, robots, IIoT intelligent devices and how the decision-making process of these machines are respecting ethical principles defined for a manufacturing facility, an industrial sector, a region, or a country.

When discussing machine ethics, the debate raises the question of how to regulate autonomous and intelligent systems-related technologies legally and the appropriate legal treatment of systems that deploy these technologies [8]. What if AI machines (e.g., intelligent IoT devices, robots, etc.,) instead of being considered people in a human sense, are put on the same legal level as corporations? It is important to remark that corporations' legal personhood can currently shield the natural persons behind them from the implications of the law [9].

As intelligent machines evolve into entities that can perceive, feel, think, and act, with intelligence comparable with animals, new regulatory and legal frameworks must be implemented to define their legal status.

9.5.10 Automation and Employment

While the concern related to AI- and automation-driven mass job losses has been a topic of concern in recent years, changes in the inherent nature of

work due to AI and automation in industrial environments will have a more substantial impact than just job losses. New forms of employment and new competencies will become the norm.

Rapid advances in AI and automation technologies can significantly disrupt labour markets. AI technology generally increases productivity in the industrial environment and, at the same time, diminish some of today's valuable employment opportunities.

The AI and automation increase and augments the productivity of some workers, and the technologies most probably replace the work done by others and likely transform all professions to some degree. The rise in automation is accelerating and occurring in a period of increasing economic inequality, fostering fears of mass technological unemployment and a reiterated call for policy efforts to address the consequences of technological change [15].

The AI effect on labour relates to automation and transforming human behaviour to assume more complex roles, departing from the physical work that dominated the industrial era repetitive administrative functions to the cognitive tasks that characterise an efficient and more productive industrial landscape.

The automation and the use of AI in industrial manufacturing can drastically cut down the human workforce, which means that revenues go to fewer persons as the wealth created by machines does not include the machines themselves. Individuals who have ownership in industrial AI-driven companies make all the money. This can widen the wealth gap, where fewer and fewer persons take a substantial portion of the economic surplus created by machines.

In this context, the ethical dilemma is connected to the occupation of humans that rely on their jobs in industry to generate income to sustain themselves and their relatives and contribute to human society.

9.6 AI and the Future Digitising Industry

The AI-based autonomous machines in industrial environments could in the future omit their traditional operating environments and increasingly move into problem areas that have earlier been available to humans due to their dynamic nature and complexity. During this transition, the intelligent machines will not be able to avoid acquiring some of the humans' limitations (e.g., learning from experience as the basis of flexible behaviour, experience as an accumulation of errors, etc.).

The development requires identifying who is responsible for the actions of machines over which humans could not have adequate control and determine ways to address the responsibility gap in moral practice and legislation [7].

9.7 Ethical Guidelines for AI in Industrial Environments

AI is increasingly impacting all industrial sectors and triggered many industrial groupings and professional bodies to provide several sets of ethical principles for AI with new ethical guidelines emerging from the British Standards Institute and the IEEE Standards Association.

The IEEE focuses on researchers' need to operate with a 'safety mindset' to pre-empt unintended or unanticipated behaviours and suggested that social and moral norms need to be considered in the design of AI technologies and applications.

The proliferation of AI-based solutions for industrial processes raises the concern that AI-related degree programmes fail to equip designers with appropriate knowledge of ethics. Related to the status of AI-based systems, IEEE [8] claims that AI should not be granted the status of "personhood", and the existing and future laws should not practically give AI legal autonomy.

As a result of globally shared needs and concerns, the industry-driven Ethics Certification Program for Autonomous and Intelligent Systems (ECPAIS) program was launched by IEEE [20]. The program's goal is to advance transparency, accountability, and reduction in algorithmic bias in autonomous and intelligent systems by setting out five core principles to consider in the design and implementation of AI and ethics: adherence to existing human rights frameworks, improving human wellbeing, ostensibly to ensure accountable and responsible design, transparent technology, and the ability to track misuse.

Another significant initiative refers to the OECD principles on Artificial Intelligence promoting innovative and trustworthy AI, respecting human rights and democratic values. The principles were adopted in May 2019 by OECD member countries when they approved the OECD Council Recommendation on Artificial Intelligence [21].

9.8 Recommendations for Ethical AI in Industrial Environments

Based on the experience presented in [14] when proposing an AI model that is intended to be retrained for downstream use, such as retraining on a new

domain or fine-tuning on a new task, the designers should inform training time and computational resources required, as well as model sensitivity to hyperparameters. This form of reporting and transparency can enable direct comparison across models, allowing subsequent users of the AI-based models to accurately assess whether the required computational resources are compatible with their setting.

The development of an industrial AI framework to enhance the explainability of AI systems is critical for all autonomous system and especially the ones that make socially significant decisions. Key to such a framework is the ability off industrial stakeholders to acquire a factual, direct, and clear explanation of the decision-making process, in the event of unwanted consequences. The specific issues addressed by different industrial sectors require the adaptation and extension of the framework to different industries.

For the evaluation of the performance of AI-based solutions for industrial applications, it is recommended to develop standardised benchmark tools, hardware-independent measurement techniques of training time (e.g., gigaflops required to convergence), and standard measurements of model sensitivity to data and hyperparameters (e.g., variance concerning hyperparameters searched).

In this context it is recommended to develop metrics for the trustworthiness of industrial AI products and services, to be used across industrial sectors. These metrics should serve as the basis for an evaluation framework that enables a user-driven benchmarking of all marketed industrial AI offerings.

The promotion of an industrial AI ethical framework must incentivise the inclusion of technical, ethical, legal, and social considerations in AI research projects and stimulate new concepts for including ethical principles into AI industrial technological developments and support the co-creation of industrial policies, standards, best practices, and rules.

AI transparency in industrial environments should be addressed from the AI system's perspective and not only from individual algorithms or components viewpoint.

AI transparency must be considered and applied concept to be interpreted in a particular context, mitigated by knowledge, information asymmetries, model-related explainability, and a set of competing interests (e.g., technological, economic). Consequently, AI transparency balances interests in industrial manufacturing processes, demanding a multidisciplinary approach that needs to be adequately addressed.

It is recommended to advance further research on computationally efficient AI algorithms, hardware, software that significantly reduce the energy consumption of AI-based solutions.

Industrial stakeholders should develop new AI technologies that advance trustworthy industrial AI to increase economic output, manufacturing efficiency, and productivity; protect natural environments; reduce emissions; and revitalise inclusive growth, sustainable development, and well-being.

They should create strategies for implementing trustworthy industrial AI across industrial sectors throughout the AI system life cycle. These include autonomy, beneficence, non-discrimination, non-bias, fairness, non-maleficence, diversity, explainability, data protection, justice, and internationally recognised labour rights.

They should implement mechanisms and safeguards, the capacity for human decisions to supervise AI-based systems.

In the following years, research and development should focus on the design of robust, secure, and safe industrial AI technologies throughout the entire life cycle in all conditions (e.g., normal use, foreseeable use or misuse, and other adverse conditions) to function appropriately and not pose any unreasonable safety risk.

The issue of industrial AI traceability should be addressed by providing mechanisms to ensure traceability (e.g., concerning datasets, processes and decisions made during the AI system lifecycle) to enable an analysis of the industrial AI system's outcomes and responses to inquiry appropriate to the industrial context.

Industrial AI stakeholders should commit to transparency and responsible disclosure regarding industrial AI systems, provide meaningful information appropriate to the industrial context to support the understanding of AI systems, make other stakeholders aware of their interactions with industrial AI systems, and understand the outcome of these systems.

Stakeholders operating in different industrial sectors should continuously develop and implement a systematic risk management approach to each phase of the industrial AI system lifecycle to address risks related to industrial AI systems.

Stakeholders designing, developing, and deploying industrial AI systems should be responsible and accountable for the proper functioning these systems and should respect the industry principles based on their roles, the industrial context, and consistent with the sector regulations and applicable laws.

Industrial actors should consider a long-term investment in research, development, and interdisciplinary activities to stimulate trustworthy AI innovation that focuses on challenging technical issues and AI-related social, legal, and ethical implications and policy issues.

In this context, it is highly recommended to foster and strengthen an interactive and collaborative European ecosystem for trustworthy industrial AI and provide mechanisms for sharing AI knowledge across industrial sectors to exchange datasets, tools, and toolchains to support the safe, fair, legal, and ethical sharing of data.

9.9 Conclusion

Digitising industry processes integrate AI-based solutions into manufacturing using autonomous learning machines based on many complex AI technologies and architectures. The digital transformation of industrial sectors thus creates new situations that call for new ethical considerations to be addressed.

It was provided a comprehensive overview of these considerations, challenges and trade-offs linked to developing AI-based intelligent stand- alone systems, IIoT systems in industrial environments as a basis for developing guidelines, standards, and norms to support their governance in industrial environments.

One such challenge relates to the question of who is responsible for the actions of an AI system?

In established industrial environments, the responsibility is attributed to a human actor, such as the owner, developer, manufacturer, or operator. However, as autonomous and learning AI-based systems become more pervasive and designed to reduce human intervention in industrial processes and accelerate automation, this may no longer be the case.

The manufacturer or operator is not always able to predict future machine behaviour, and thus in specific cases cannot be held responsible. This calls for new regulations to be in place to support decisions related to who is accountable or faces a responsibility gap that traditional concepts cannot bridge.

This article pases awareness of diverse and complex ethical concerns arising from the deployment of AI in industrial environments: from the degradation of the environment to job losses due to automation. These concerns may differ in interpretation, focus, and weight within various industries and organisations, mainly because ethical terminology, principles, and approaches – although necessarily aligned to society's common and

recognised values – will vary to adapt to the ecosystem in each industry or organisation. Consequently, no one solution can mitigate all concerns, so this article aims to spark new topics for further research.

Digitising industry processes integrate into the manufacturing processes AI-based solutions using autonomous learning machines based on neural networks, genetic algorithms, and agent architectures. The digital transformation of industrial sectors creates a new situation.

Acknowledgements

Part of the work presented in this chapter was supported by the European Commission within the European Union's Horizon 2020 research and innovation programme funding, ECSEL Joint Undertaking project AI4DI under Grant Agreement No. 826060, ECSEL Joint Undertaking project ArchitectECA2030 under Grant Agreement No. 877539, ECSEL and ECSEL Joint Undertaking project AI4CSM under Grant Agreement No. 101007326.

References

[1] AI4DI (2019). Artificial Intelligence for Digitising Industry. Available at: https://ai4di.eu/.

[2] European Commission (2018). Communication from the Commission to the European Parliament, the European Council, the Council, the European Economic and Social Committee and the Committee of the Regions on Artificial Intelligence for Europe, Brussels, 25.4.2018 COM (2018) 237 final. Available online at: https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:52018DC0237&from=en

[3] European Commission (2020). On Artificial Intelligence - A European approach to excellence and trust. White Paper. Available online at: https://ec.europa.eu/info/sites/default/files/commission-white-paper-artificial-intelligence-feb2020_en.pdf

[4] European Commission - High-Level Expert Group on Artificial Intelligence (2019). Ethics guidelines for trustworthy AI. Available online at: https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

[5] European Commission (2021). Commission staff working document. Impact assessment. Accompanying the proposal for a regulation of the European Parliament and of the Council. Laying down harmonised

rules on artificial intelligence (Artificial Intelligence Act) and amending certain union legislative act. Available online at: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A52021SC0084

[6] N. J. Nilsson, (2010). The Quest for Artificial Intelligence: A History of Ideas and Achievements, Cambridge, UK Cambridge University Press.

[7] A.Matthias, (2004). The responsibility gap: Ascribing responsibility for the actions of learning automata. Ethics and Information Technology, Sept 2004, Vol. 6, Issue 3, 175-183. Available online at: https://link.springer.com/content/pdf/10.1007/s10676-004-3422-1.pdf

[8] IEEE - The Institute of Electrical and Electronics Engineers (2017). Ethically Aligned Design: First Edition. A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems. Available online at: https://standards.ieee.org/content/dam/ieee-standards/standards/web/documents/other/ead_v2.pdf

[9] E., Bird, J., Fox-Skelly, N., Jenner, R., Larbey, E.. Weitkamp and A., Winfield, (2020). The ethics of artificial intelligence: Issues and initiatives. Available online at: https://www.europarl.europa.eu/RegData/etudes/STUD/2020/634452/EPRS_STU(2020)634452_EN.pdf

[10] W., Knight, (2020). AI Can Do Great Things—if It Doesn't Burn the Planet. Available online at: https://www.wired.com/story/ai-great-things-burn-planet/

[11] S., Meinecke, (2018). AI could help us protect the environment — or destroy it. Available online at: https://www.dw.com/en/ai-could-help-us-protect-the-environment-or-destroy-it/a-44694471

[12] D., Lu, (2019). Creating an AI can be five times worse for the planet than a car. Available online at: https://www.newscientist.com/article/2205779-creating-an-ai-can-be-five-times-worse-for-the-planet-than-a-car/

[13] ML CO2 Impact. Available online at: https://mlco2.github.io/impact/#home

[14] E., Strubell, A., Ganesh, A., McCallum, (2019). Energy and Policy Considerations for Deep Learning in NLP. Available online at: https://arxiv.org/pdf/1906.02243.pdf

[15] M. R., Frank, et al., (2019). Toward understanding the impact of artificial intelligence on labor. Available online at: https://www.pnas.org/content/pnas/116/14/6531.full.pdf

[16] Y., Wang, X., Mengran, H. G. T. Olya, (2020). Toward an Understanding of Responsible Artificial Intelligence Practices. In Proceedings of the 53*rd* Hawaii International Conference on System Sciences. Available

online at: https://scholarspace.manoa.hawaii.edu/bitstream/10125/64352/0491.pdf

[17] WEF-World Economic Forum (2018). Harnessing Artificial Intelligence for the Earth. Available online at: http://www3.weforum.org/docs/Harnessing_Artificial_Intelligence_for_the_Earth_report_2018.pdf

[18] NIST (2019). US Leadership in AI: A Plan for Federal Engagement in Developing Technical Standards and Related Tools. Washington: NIST (US Department of Commerce), 8. Available online at: https://www.nist.gov/system/files/documents/2019/08/10/ai_standards_fedengagement_plan_9aug2019.pdf

[19] Patrizio, A. (2018). IDC: Expect 175 zettabytes of data worldwide by 2025. Available online at: https://www.networkworld.com/article/3325397/idc-expect-175-zettabytes-of-data-worldwide-by-2025.html#:~:text=By%202025%2C%20IDC%20says%20worldwide,cloud%20as%20in%20data%20centers.&text=IDC%20has%20released%20a%20report,study%2C%20the%20numbers%20are%20staggering.

[20] The Ethics Certification Program for Autonomous and Intelligent Systems (ECPAIS). Available online at: https://standards.ieee.org/industry-connections/ecpais.html

[21] Organization for Economic Co-operation and Development (2019). Principles on Artificial Intelligence. Paris: OECD. Available online at: https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449

[22] Imperial College London. (2017). Written Submission to House of Lords Select Committee on Artificial Intelligence [AIC0214]. Available online at: http://bit.ly/2yleuET

[23] T., King, N., Aggarwal, M., Taddeo, and L. Floridi, (2018). Artificial Intelligence Crime: An Interdisciplinary Analysis of Foreseeable Threats and Solutions. Available online at: https://ssrn.com/abstract=3183238

[24] Montreal Declaration for a Responsible Development of Artificial Intelligence. (2017). Announced at the conclusion of the Forum on the Socially Responsible Development of AI. Available online at: https://www.montrealdeclaration-responsibleai.com/the-declaration

[25] Partnership on AI. Safety Critical AI. Tenets. Available online at: https://partnershiponai.org/program/safety-critical-ai-2/

[26] Asilomar AI Principles. (2017). Principles developed in conjunction with the 2017 Asilomar conference [Benevolent AI 2017]. Available online at: https://futureoflife.org/ai-principles

[27] J., Cowls, and L. Floridi (2018). Prolegomena to a White Paper on Recommendations for the Ethics of AI (June 19, 2018). Available online at: https://ssrn.com/abstract=3198732

[28] House of Lords Artificial Intelligence Committee. (2018). AI in the UK: ready, willing and able? Available online at: https://publications.parliament.uk/pa/ld201719/ldselect/ldai/100/10002.htm

[29] European Group on Ethics in Science and New Technologies. (2018). Statement on Artificial Intelligence, Robotics, and "Autonomous" Systems. Available online at: https://op.europa.eu/en/publication-detail/-/publication/dfebe62e-4ce9-11e8-be1d-01aa75ed71a1

[30] ArchitectECA2030 (2020). Trustable Architectures with Acceptable Residual Risk for the Electric, Connected and Automated Cars. Available at: https://autoc3rt.automotive.oth-aw.de/

[31] AI4CSM (2021). Automotive Intelligence for Connected Shared Mobility. Available at: https://ai4csm.automotive.oth-aw.de/