IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 28 January 2024, accepted 2 April 2024, date of publication 1 May 2024, date of current version 31 May 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3395444

TOPICAL REVIEW

Explainable AI for Cyber-Physical Systems: Issues and Challenges

AMBER HOENIG[1], KAUSHIK ROY[2], YAA TAKYIWAA ACQUAAH[2], SUN YI[3], AND SALIL S. DESAI[4,5]

[1]Department of Computational Data Science and Engineering, North Carolina Agricultural and Technical State University, Greensboro, NC 27411, USA
[2]Department of Computer Science, North Carolina Agricultural and Technical State University, Greensboro, NC 27411, USA
[3]Department of Mechanical Engineering, North Carolina Agricultural and Technical State University, Greensboro, NC 27411, USA
[4]Department of Industrial and Systems Engineering, North Carolina Agricultural and Technical State University, Greensboro, NC 27411, USA
[5]Center of Excellence in Product Design and Advanced Manufacturing, North Carolina Agricultural and Technical State University, Greensboro, NC 27411, USA

Corresponding author: Amber Hoenig (ashoenig@aggies.ncat.edu)

This research is based upon work supported by the Office of Naval Research (ONR) (Award No. N00014-22-1-2724). The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the U.S. Government.

ABSTRACT Artificial intelligence and cyber-physical systems (CPS) are two of the key technologies of the future that are enabling major global shifts. However, most of the current implementations of AI in CPS are not explainable, which creates serious problems in ethical, legal, regulatory, and other domains. Therefore, it is necessary for explainable artificial intelligence (XAI) to be integrated with cyber-physical systems to meet the vital needs for control, fairness, accountability, safety, cyber-resilience, and cybersecurity. The goal of this review is to demonstrate the need, benefits, challenges, and implementation of XAI for CPS. We review the existing literature about XAI and CPS, discuss the current state of the art, examine applications in different domains, and make recommendations for future research directions. To the best of our knowledge, this is the first peer-reviewed academic article to provide a comprehensive review of general XAI for CPS. We also contribute new research ideas including development of multisensory explanations and outputs for these systems, application of XAI to CPS to decrease occupational burnout and increase employee engagement, and enumeration of the multidisciplinary goals and benefits of XAI as applied to cyber-physical systems.

INDEX TERMS Cyber-physical systems (CPS), cyber-resilience, cybersecurity, explainable artificial intelligence (XAI), industrial CPS, Industry 5.0.

I. INTRODUCTION

Explainable artificial intelligence (XAI) is the set of methods used to provide clear and understandable explanations of how an AI model works, either at the level of a single instance or the entire system [1]. Cyber-physical systems (CPS) are systems which integrate physical, sensor, and cyber components for the purpose of monitoring and control of elements in the physical world [2], [3], [4]. Together, XAI and CPS create a foundation for cyber-physical systems that are secure, accountable, trustworthy, ethical, flexible, and cyber-resilient.

The world's growing dependence on artificial intelligence (AI) is evident in its widespread use across various applications such as natural language processing (NLP), data mining, and social media. As technology advances, AI continues to play an essential role in shaping diverse industries. CPS are one of these critical areas of adoption of AI, allowing the systems to gain the benefits of automated machine learning (ML) and intelligence. The implementation of XAI with high technology and CPS has the potential to generate pivotal changes in human society and the environment [5], creating ways to reduce resource usage, build collaborative relationships between human workers and machines, increase safety, improve cybersecurity, develop cyber-resilience, achieve goals for future industry development, and implement socially and environmentally sustainable AI. This is a time of intense, pervasive global shifts in industry and life driven by changes in technology and AI.

As more processes become automated and integrated into CPS, the need for optimization, security, cyber-resilience,

The associate editor coordinating the review of this manuscript and approving it for publication was Moussa Ayyash.

and safety grows. XAI is needed in these systems to ensure that they meet these goals as well as maintaining fairness, proper use of AI, and ethical and legal compliance. XAI also ensures better user understanding of CPS, enhanced human-machine collaboration, decreased use of resources, and lower costs. Many complicated physical processes are becoming automated as global industry surges forward. The resulting CPS have high levels of complexity in their components and behavior [6]. Such complexity necessitates clear explainability to enable understanding and control of the system.

However, several challenges presently impede the development of optimal XAI for CPS. Few articles exist in the literature on the topic of XAI and general CPS. The great majority of the available articles focus on a specific topic such as XAI for medical CPS [7], XAI applied to industrial CPS [8], or XAI used in CPS security [9], [10]. There is a need for quantitative analysis of the numerous benefits of using XAI with CPS, and there is a lack of standardized methods of quantifying the understandability and performance of XAI. Existing XAI models still have the possibility of containing bias, especially if they are trained on manually annotated or otherwise biased data. Because of the complex nature of the field and the multitude of applications, interdisciplinary collaboration is urgently needed [5], [11]. Furthermore, information dissemination and knowledge transfer from meaningful XAI explanations and interfaces will allow CPS and their related fields to progress rapidly.

Explainability increases trust and the ability to properly implement, manage, monitor, and alter the system as needed. XAI makes it possible to generate explanations for CPS, but most existing XAI methods are not sufficient. Most of the currently used XAI methods do not adequately address time-series information and are not context-aware [6]. In addition, the explanations given often fail to provide true understandability or meaningful visualizations. XAI methods applied to CPS need to be accurate, provide clear explanations, integrate contextual information, create meaningful visualizations to convey information, and provide customized explanations for different types of users [6].

In view of these challenges, this article presents a comprehensive review and analysis of XAI for CPS. We demonstrate that integrating XAI into cyber-physical systems solves many of the current problems facing AI-based CPS, including cyber-resilience, fairness, ethical and legal compliance, transparency, safety, cybersecurity, control, and many others. We assess the current research landscape, discuss current challenges, demonstrate the necessity of implementation, and propose directions for future research with the aim of providing solutions for potential barriers and displaying the benefits of implementation of XAI for CPS. The motivation behind this review stems from the increasing importance of AI in CPS, where trust, reliability, and accountability are paramount. Transparency and understandability are vital if stakeholders are to implement these systems, for without understanding, the stakeholders may not trust the system and its decisions, which can prevent them from integrating much-needed AI into their cyber-physical systems. The need for understandable, resilient, and safe AI becomes even more apparent as safety-critical CPS are implemented in areas such as healthcare, self-driving vehicles, and life-sustaining infrastructure. Our goal is to demonstrate the manifold benefits of applying XAI to CPS, remove barriers to its application, and provide guidance for future research to create optimal implementations of XAI for CPS. To our best knowledge, this is the first peer-reviewed academic article to provide a comprehensive review of general XAI for CPS. We not only provide detailed information about requirements, existing implementations, numerous critical areas of application, and current gaps; we also contribute new recommendations for future researchers including development of multisensory outputs, use in occupational burnout reduction, and facilitation of employee engagement.

The scope of the review encompasses the implementation and applications of XAI for cyber-physical systems. We examine the challenges which prevent the application of AI for CPS, demonstrate the need for explainability in these systems, and delineate the many benefits of integrating XAI into CPS. We demonstrate how XAI can be integrated into CPS to meet Industry 4.0 and Industry 5.0 goals. We also review existing implementations of XAI for CPS, reveal current challenges, and recommend potential solutions and improvements.

This article is organized as follows. In Section II, we provide the background of explainable artificial intelligence, enumerate its key goals and benefits, discuss current applications of AI, and review the requirements of XAI models. We delineate ways in which XAI can be used to manage AI and then discuss the risk of bias in traditional AI models. In Section III, we define and discuss cyber-physical systems, explain the role of CPS in Industry 4.0, demonstrate the roles CPS play in modern critical infrastructure, discuss the importance of cybersecurity and cyber-resilience in CPS, and illustrate the risks of unexplainable AI in CPS.

Section IV provides a detailed discussion of XAI in CPS. We discuss barriers that hinder adoption of AI in cyber-physical systems and propose XAI as a solution to these issues. We explore the connections between AI, XAI, and CPS in Industry 4.0 applications. Next, we provide a detailed discussion of related research about XAI for cyber-physical systems. We then review challenges in current implementations of XAI for CPS and offer recommendations to mitigate them. The next subsections review XAI and CPS in systems engineering, the dangers of bias in CPS, XAI in medical CPS, and the importance of understandability in CPS. Finally, we provide a requirements analysis for development of effective XAI-based cyber-physical systems.

In Section V, we explore current challenges in the domain of XAI for CPS. We discuss the importance of XAI in future CPS, the challenges in current implementations, and legal and ethical concerns. We review important issues and

vulnerabilities in medical cyber-physical systems; industrial, smart city, and environmental applications; CPS-based critical infrastructure; and cybersecurity in CPS.

Section VI provides recommendations to solve existing gaps in this domain, highlighting the lack of formalized standards, the need for expert supervision, the importance of customized user outputs, the potential for multisensory explanations, and the urgent need for interdisciplinary collaboration and communication. In Section VII, we provide future research directions and recommendations including applications to Industry 5.0, prevention of occupational burnout, multiple CPS integration, system cognition, self-explainability, and context awareness.

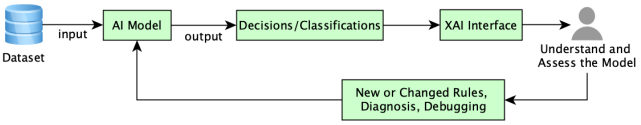

FIGURE 1. The basic process of XAI.

II. EXPLAINABLE ARTIFICIAL INTELLIGENCE

A. BACKGROUND OF XAI

Explainable artificial intelligence is the system of methods used to provide human-understandable explanations of the rationale behind decisions or recommendations made by AI-based systems. This can be at the local level, explaining a specific data instance or prediction, or at the global level, explaining the entire model [1]. According to Rai [12], "Explainable AI (XAI) is the class of systems that provide visibility into how an AI system makes decisions and predictions and executes its actions. XAI explains the rationale for the decision-making process, surfaces the strengths and weaknesses of the process, and provides a sense of how the system will behave in the future." XAI includes algorithms, tools, and approaches designed to create these explanations for AI [13]. It uses visualization and natural language descriptions to communicate the reasoning behind decisions made by an AI system [14]. Once users understand that reasoning, they can use the information to modify, debug, or diagnose the system. XAI can also show users existing policies learned by an algorithm, which the users can then analyze, interpret, and change [15]. Even specific kinds of concerns can be addressed through this type of XAI without having to retrain the system. Fig. 1 demonstrates the basic process of explainable AI. Data representations are input into the AI model, which outputs decisions or classifications for each instance. The XAI interface displays explanations about the logic used to make decisions in the AI model. This is provided to the human user, who can use this information to understand and assess the model, then provide new or changed rules, diagnosis, or debugging to the AI model.

XAI can be traced back to the 1970s, yet it only recently came into focus with the modern growth of deep learning (DL) and as concerns have been raised about the legality, ethics, bias, and unethical uses of AI systems [16], [17]. As XAI becomes more widespread, implemented in cyber-physical systems such as smart agriculture, disaster response and prevention, smart cities, smart power grids, intelligent transportation, and beyond, affecting resource usage, safety, and the environment, it has the potential to affect all life on earth. Artificial intelligence is already applied in environmental, sociopolitical, financial, supply chain, production, industrial, chemical, and many other applications. XAI has already been implemented in flood susceptibility mapping [18], drought prediction [19], wildfire susceptibility analysis [20], suitable crop prediction based on soil and environmental conditions [21], and many other environmental applications. Implementation of XAI into cyber-physical systems can promote safety, reduce bias, promote fairness, decrease waste, save money, prevent hazardous materials accidents, and reduce resource usage (Table 1).

TABLE 1. Key goals of explainable artificial intelligence by domain.

	Key Goals and Benefits of XAI
General	Communicate and foster understanding of the process within AI models, explain decision making, build trust, understand probable future behavior of the model, find strengths and weaknesses of the model, diagnosis, debugging, improvement, use of clearer understanding of the model to improve security, simplify process of KDD (Knowledge and Data Discovery), apply or add desired rules to improve or change the model without requiring the model to be entirely regenerated, enable human monitoring and intervention with AI models
Conceptual Applications	Ethics, transparency, legal compliance, bias detection, bias reduction, trust, fairness, control and management of AI, cybersecurity improvement
Cyber-Physical Systems	Improve control of CPS, increase understanding of cyber-physical system functions and components, enhance cybersecurity, increase stakeholder willingness to implement AI by building trust, improve efficiency, decrease costs, decrease losses, teach users about relevant features in decision making, provide explanations at global and local levels during inquiries, verification and validation of AI-CPS processes, documentation and justification of decisions, regulatory compliance, legal accountability, preparedness and optimization for varying conditions, flexibility, cyber-resilience, safety
Cyber-Physical Systems and Human Factors	Enable creativity, facilitate engagement, create sustainable working conditions, educate users, enable human-machine collaboration, meet Industry 5.0 goals for social sustainability in industry, improve and broaden human supervision over systems
Industrial and Cyber-Physical Systems Applications	Decrease waste, decrease use of resources, increase stakeholder trust to enable implementation of AI, increase efficiency, facilitate operational management via data analysis, enable analysis and prediction of smart machine operations, assist in the systems engineering process, generate key performance indicators (KPIs), generate predictive maintenance recommendations, allow product customization, decrease costs, prevent accidents, create cyber-resilience
Environmental	Decrease waste, decrease resource usage, monitor environment, make recommendations to protect environment, agricultural applications, accident prevention in hazardous materials transportation, smart cities, smart power grids, water management, flood detection
Scientific and Research Directions	Interdisciplinary communication and collaboration, expert supervision and learning, discovery of predictive features, effective and testable big data analysis, creation of understandable explanations for people of different cultures/fields of study/levels of expertise, increase stakeholder trust to enable implementation of AI, improvement of AI systems, diagnosis and modification of AI systems

B. AI APPLICATIONS AND THE BLACK BOX PROBLEM

Artificial intelligence has revolutionized much of the modern world. An ever-increasing number of AI-based applications have become integrated with cyber-physical systems such as cyber-physical manufacturing systems, chemical engineering, additive manufacturing, biomanufacturing, and nanomanufacturing [5], [22], [23], [24]. AI is used in robotics, health information technology, intrusion detection systems, social media, chatbots, predictive policing, in-store retail technology, and mobile applications [12], [14], [25]. Rai [12] states that AI "systems are penetrating a broad range of industries, such as education, construction, healthcare, news and entertainment, travel and hospitality, logistics, manufacturing, law enforcement, and finance.' Artificial intelligence affects who we become friends with, how we care for our children and elders, our purchasing and hiring decisions, what we like, our opinions, and the newsfeeds we are shown [12], [26]. AI is used in high-stakes applications affecting human life, including determining credit approval, operation of automated defense machines, in biomedical fields, and in cybersecurity [7], [16]. However, most of these AI applications are not explainable.

There is often a tradeoff between the inherent explainability of a model and its prediction accuracy. Simple models are easier to understand than complex, many-layered models that implement extensive mathematical operations, but simple models lack the predictive power of more complicated models such as deep learning [27]. Deep learning models are widely used because they can analyze very large datasets and provide high result accuracy [28]. These models are noted for their ability to find meaningful correlations between variables, no matter how seemingly small they may be [27].

In general, the more complex a type of model is, the better its performance will be [29], although some researchers such as Barredo-Arrieta et al. [30] and Speith [31] disagree. The use of highly complicated models leads to the black box problem: the more complex a model is, the more difficult it is to understand what is going on inside it [29]. Many deep learning models are so complex that it is impossible for a human to monitor every equation and connection that leads to the final output. Often, only the input and output are visible [27]. XAI solves the black box problem. It is used to create AI models which are high-performing, yet still provide a means of being human-understandable [1]. This enables trust, understanding, and effective management of those models.

C. DEFINITION AND REQUIREMENTS OF XAI MODELS

XAI can be defined as the subfield of artificial intelligence which "aims to develop more understandable models while maintaining a high degree of learning performance (prediction accuracy); and enable human users to comprehend, adequately trust, and manage the future generation of artificially intelligent partners" [32]. The goal of explainable artificial intelligence is "to provide reasoning for ML model outputs, allowing humans to understand and trust ML models" decision-making process" [1]. Thus, XAI is the subset of AI models which are understandable by humans. This includes new AI models that are human-understandable, existing inherently understandable models, or existing AI models which are modified to become human-understandable [1].

XAI models may be inherently interpretable (also called transparent) or uninterpretable (black-box or opaque) [12], [32], shown in Table 2. Generally, inherently interpretable models are simpler and smaller in both size and number of node connections. Opaque models such as deep learning models are often so complex that even experts often cannot give a clear explanation as to what is happening within them [31]. XAI models may also be local, providing an explanation for a single instance or output, or global, giving an explanation for the behavior of the entire model. Transparent XAI models are also called ante-hoc as the model itself is inherently explainable. Post-hoc XAI models are used to explain a model after it has been trained. Typically, post-hoc methods are used to explain opaque AI models [31]. Another way to define XAI models is by whether they are model-specific (explanatory methods used to explain a certain type of model) or model-agnostic (explanatory methods that can be applied independent of the model type) [33].

XAI models must meet the requirements of demonstrating fidelity as well as being comprehensible, transparent, explainable, understandable, and accurate [7], [34].

- **Comprehensibility** or **interpretability** is defined as "the capacity of a learning system to express its learned information in a human-comprehensible manner" [7], [34].
- A model is **transparent** if it is comprehensible by itself without any additional outside information needed [34].
- If a model is **explainable**, it acts as an interface between people and the AI program making decisions or recommendations. An explainable model must give a clear understanding to human users while maintaining a close approximation of what is going on within the AI program to produce the decisions [30], [34].
- A model is considered to be **understandable** if a human is able to understand what the model is doing without having to understand the details of the inner workings of the model or its algorithmic structures [30], [34]. Ideally, descriptions need to be given in relatively small sections to facilitate understanding. These descriptions should provide a clear connection between qualitative and quantitative concepts [34].
- **Accuracy**: XAI models must demonstrate the ability to correctly predict unseen instances [7]. This can be quantified with evaluation measures such as accuracy, F1 score, and others.
- **Fidelity**: The ability to accurately represent or imitate what is going on inside a black box model [7]. Fidelity is measured similarly to accuracy, but instead of comparing the predictions to the correct outcomes, the

predictions are compared to the outcomes of the black box model.

It is important that the output of the XAI is well suited to and understandable by the individuals who will be using it [7]. This particularly necessary for users from non-machine-learning-related fields.

D. XAI FOR MANAGEMENT OF AI

XAI is not just needed for understanding AI; it is critical for managing it [35]. XAI is used to explain and justify decisions made by AI models, control and improve AI models, and explore relationships within data [36]. AI has changed the way humans live, interact, spend their money, and even vote [12], [35]. It has changed what people believe through algorithmic bias, curated information feeds in online news and social media, and AI-propagated misinformation and disinformation. Society as a whole is affected by AI, from the level of individuals to entire governments [35]. XAI is a vital tool that can be used to manage and help control the effects of artificial intelligence. It can allow lawmakers and policymakers to understand the artificial intelligence being used, make appropriate laws, and be better able to enforce them. XAI also makes it possible to provide meaningful explanations to citizens about AI-based decisions made on their behalf without releasing intellectual property [27].

E. BIAS IN AI

Existing AI algorithms are prone to bias. Artificial intelligence algorithms function and make predictions based on the data they were trained on. They learn connections, characteristics, and classifications based on their training dataset. If that dataset was biased, all the decisions the AI network makes will be affected by that bias. In other words, the parameters learned by the network and thus the decisions it makes will ''represent truth in a training data set, rather than truth in the world'' [27]. Algorithms that were trained on a dataset which is biased against certain groups will then make predictions biased against that group. Biased training datasets will result in unfair and inaccurate decisions that affect underrepresented or minority groups. Machine learning models are prone to repeating existing systemic bias from human decision-makers [25]. Biases in artificial intelligence algorithms ''have led to large-scale discrimination based on race and gender in a number of domains ranging from hiring to promotions and advertising to criminal justice to healthcare'' [12]. Some countries use AI with cyber-physical systems such as street cameras for crime detection and prevention [5]. AI is used in facial recognition for police operations and for recidivism prediction in the field of criminal justice [33]. If the training data are biased, this can lead to biased and unjust results.

III. CYBER-PHYSICAL SYSTEMS

A. BACKGROUND OF CYBER-PHYSICAL SYSTEMS

According to Alguliyev et al. [2], a cyber-physical system can be defined as ''a system that can effectively integrate cyber and physical components using the modern sensor, computing and network technologies.'' Lee et al. [4] define cyber-physical systems as ''transformative technologies for managing interconnected systems between its physical assets and computational capabilities.'' Cyber-physical systems are used to intelligently monitor and control elements in the physical world [3]. CPS integrate sensors, control systems, and embedded systems. The block diagram shown in Fig. 2 displays the components and connectivity of a cyber-physical system. The diagram also includes external network and storage components, which are often included in cyber-physical systems.

From critical infrastructure to innovative smart devices, cyber-physical systems have become a vital part of everyday functioning for people and systems worldwide [9]. Cyber-physical systems are becoming more complex, ubiquitous, and deeply integrated with daily life [29]. Modern critical infrastructure is heavily dependent on cyber-physical systems. CPS offer tremendous benefits in terms of economics, efficiency, information flow, production planning, decreased downtimes, improved product quality, and cost reduction [1]. Critical areas of interest in the development of cyber-physical systems include reliability, security, and resilience [1].

Data can be used as an engineering tool in development of future CPS infrastructure [37]. Data analysis can be applied not only in the physical or cyber structure being engineered, but also throughout the entire system life cycle to monitor

TABLE 2. Inherently interpretable and uninterpretable artificial intelligence models [12], [30], [31].

	Inherently Interpretable Models	Uninterpretable/Black Box Models
Examples	• Spare linear models • Decision trees • K-nearest neighbors • Bayesian classifiers • General additive models • Rule-based learners • Case-based reasoning	• Recurrent Neural Networks • Convolutional Neural Networks • Long Short-Term Memory Networks • Multiple Classifier (Ensemble) Systems
Characteristics	• Components can be directly inspected • Easier to understand how predictions are made by the model • Lower number of internal components (e.g., rules, features, nodes, paths) • Transparent, traceable, and interpretable • Can lose inherent interpretability if too many rules or parameters are included • Typically lower accuracy than black box models (see counterarguments in [30], [31]) • Relatively simple structure	• Direct inspection of components is not inherently meaningful • Difficult to understand how predictions are made by the model • High number of internal components (e.g., nodes, connections, paths, rules, features) • Less transparency, traceability, and interpretability • Complicated structure • Typically higher accuracy than inherently interpretable models • Inputs and outputs have highly nonlinear associations

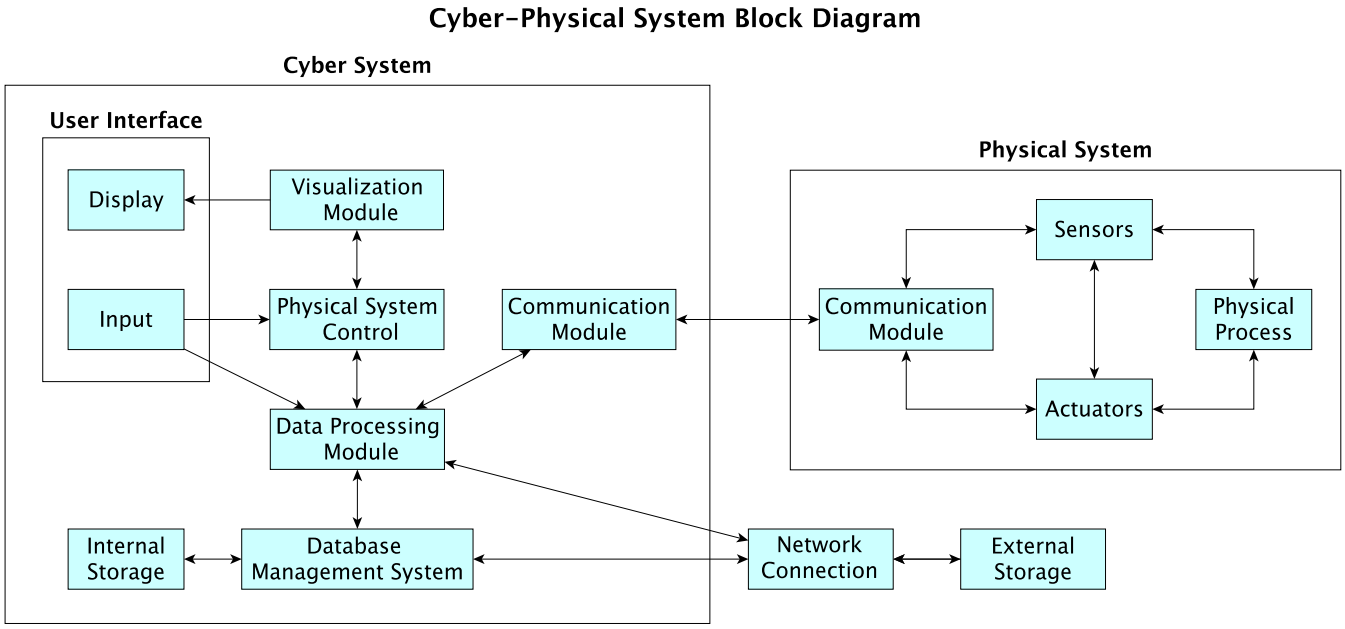

FIGURE 2. A block diagram showing the components of a cyber-physical system.

decisions, make predictions, make decision recommendations, and learn from existing information. Using data to design, implement, and integrate multiple systems, such as transport systems, railways, and bridges, can lead to more sustainable infrastructure and decisions [37].

B. CYBER-PHYSICAL SYSTEMS IN INDUSTRY 4.0

Cyber-physical systems comprise one of the foundational technologies of the Fourth Industrial Revolution (Industry 4.0). Industry 4.0 "forms the process of combining technologies and knowledge, providing autonomy, reliability, systematicity, and control without human participation" [2]. According to Zizic et al. [38], "Industry 4.0 is based on the concept of smart factory, where smart products, machines, storage systems, and data unite in the form of the cyber-physical production systems."

Technologies integrated in Industry 4.0 include cloud computing, the Internet of Things (IoT), smart technologies, and big data [2]. Cyber-physical systems can be considered as the key element in the Fourth Industrial Revolution [2], [39]. The goal of Industry 4.0 is to combine physical components and technology to create outstanding operational performance and productivity [13]. Automation, interoperability, and process development are further goals of this movement.

In short, Industry 4.0 integrates computerized production systems, data, and information systems [7]. It fuses technologies between physical, digital, and biological domains [39]. The key AI areas within the Fourth Industrial Revolution include machine learning, deep learning, natural language processing, data science, and computer vision [13]. Industry 4.0 integrates cyber-physical systems, robotics, additive manufacturing, big data, augmented reality, IoT, system integration, and cloud computing. Many applications of Industry 4.0 involve cyber-physical production systems, integrating physical and digital systems so they are synergistically involved in all stages of product development [40].

Fundamental cyber-physical system types in Industry 4.0 include smart factories, chatbots, human-computer interaction, smart healthcare, manufacturing, smart products, augmented reality, transportation industry, aviation, autonomous vehicles, smart consumer appliances, intelligent chemical industry, industrial robots, smart assistance, autonomous resource exploration, cybersecurity and privacy, predictive maintenance, and smart cities [2], [5], [13], [41]. Smart cities may employ diverse CPS such as smart transportation, smart airports, smart ports, smart hospitals, flood detection and mitigation, drainage monitoring systems, and smart power grids [34], [42].

C. CYBER-PHYSICAL SYSTEMS AND MODERN INFRASTRUCTURE

Cyber-physical systems are the basis for much of the modern infrastructure of the world. Examples of cyber-physical systems include smart medicine, defense systems, smart grids, modern meteorology, disaster prevention and response, emergency response systems, smart cities, autonomous vehicles, mobile communication systems, smart manufacturing, wearable smart devices, smart agriculture, smart medical service systems, smart buildings, factory automation, building and environmental control, smart infrastructures, process control, assistive devices, smart homes, smart transportation, intelligent highways, robotic systems, and aerospace systems [2], [7], [11], [29], [43], [44]. Some researchers have even proposed the development of planetary-scale cyber-physical systems [11].

With this explosive growth in CPS technology and its widespread implementation come great risks in confidentiality and security [2]. As humans rely more heavily on large-scale cyber-physical systems for their daily needs, those cyber-physical systems become targets for those who wish to cause harm. If water resource management systems, food supply, power grids, and other life-sustaining resources are connected as cyber-physical systems, they become vulnerable to attacks which could have widespread consequences for human life [45]. Cybercrime can cause great damage on a global scale. It affects society at all levels, including individuals and groups, industry, educational institutions, businesses, national security, and international security [46].

Advanced persistent threats (APT) are particularly dangerous to critical infrastructure and the economy, challenging existing traffic detection technology [47]. Original attack samples usually are not shared by entities which have been attacked due to concerns of embarrassment and data privacy; therefore, knowledge of new and evolving APT attack methods is limited [47]. It is therefore important for any developed intrusion detection systems (IDS) for industrial systems to be well equipped to detect and prevent new attacks; harnessing the power of big data analysis, cloud-based computing, and deep learning has helped in the pursuit of these goals. The use of explainable AI with APT edge defense models can raise ''the level of protection and defense against APTs at the edge'' [47].

D. CYBERSECURITY AND CYBER-RESILIENCE IN CPS

One of the most important facets of cyber-physical systems is cybersecurity [2]. As CPS is applied to many different domains, systems become newly vulnerable to cyberattacks; this threatens data confidentiality, trade secrets, and infrastructure. These vulnerabilities ''include cyberattacks via internet-connected devices, and physical assaults which can lead to supply chain disruption or system failures'' [9]. The massive amount of data being produced, transferred, and stored in cyber-physical systems makes them a potential goldmine for hackers, data aggregators, advertisers, governments, and other organizations. In addition, the increasing reliance on cyber-physical systems means that any attacks that prevent the systems from functioning could have widespread consequences such as medical facilities' inability to access patient medical records, communication problems, manufacturing facilities shutting down, and widespread political and economic consequences. Cyber-resilience is therefore a crucial element in CPS, allowing the systems to continue to function as intended in spite of cyberattacks.

Industrial CPS are particularly susceptible to cyberattacks with potentially dangerous consequences. According to Alqaralleh et al. [41], ''Cyber-attacks on cyber-physical systems (CPS) [result in] sensing and actuation misbehavior, severe damage to physical object, and safety risk.'' One type of commonly used attack against industrial manufacturing CPS is adaptive poisoning. In an example given by Li et al. [48], the ''attacker can manipulate the [federated learning] model by injecting poisoned data into each training epoch,'' thus training the model incorrectly so it will yield wrong results. Such attacks are becoming more difficult to detect and thus prevent as methods with greater concealment are developed. Li et al. [48] propose a multitentacle federated learning method to both detect and minimize damages from adaptive poisoning attacks in such networks.

Another serious threat to industrial CPS is adversarial example. Li et al. [49] state that adversarial example is ''gradually growing into the greatest threat to deep learning.'' Its purpose is not to steal private information but to deceive the DL model through false data. In industrial CPS, this can lead to serious accidents as well as financial and economic losses. Even ultra-low frequency insertion of adversarial examples into a DL model can lead to serious damage. The authors propose a Decentralized Swift Vigilance (DeSVig) framework to ''circumvent unknown adversarial examples in industrial artificial intelligence systems'' [49]. This model enables industrial AI systems to rapidly recognize and correct abnormal inputs within seconds.

One of the recent and growing developments in the field of machine learning is generative adversarial networks (GANs). GANs can be used to generate or to detect false data; GANs pose a risk for being used in malicious attacks such as insertion of false data into systems (e.g., data poisoning attacks), but they can also be harnessed to protect against adversarial attacks by improving IDS. Generative adversarial networks can be defined as ''a system of two neural networks (generator and discriminator) that compete with each other in a kind of finite zero-sum game'' [40]. The goal of the generator is to create instances that closely match a given training set, while the goal of the discriminator is to determine which of the instances were generated and which belong to the training set. Both networks are trained simultaneously. Different methods of adversarial example generation have been created [49]. GANs can be used in both creation and detection of adversarial attacks. For example, GANs can generate samples for use in attacks on federated learning models, yielding greater attack efficiency [48]; GANs can also be combined with CNNs for accurate intrusion detection [47].

Marino et al. [14] state of adversarial machine learning, ''Adversarial samples are crafted from an attacker perspective to evade detection, confuse the classifier, degrade performance and/or gain information about the model or the dataset used to train the model. Adversarial samples are also useful from a defender point of view given that they can be used to perform vulnerability assessment, study the robustness against noise, improve generalization and debug the machine learning model.'' Any XAI-based explanations of these findings would provide further insight into the results such as what makes the system vulnerable, what features are most likely to be involved in misclassified instances or failed attack detection, or what characteristics create robustness. For instance, Marino et al. [14] used adversarial learning to

create an explainable method to identify the most relevant features in misclassified instances in an IDS by calculating the difference between the false generated samples and the original real samples. Barredo-Arrieta et al. [30] note that, ''Once trained, generative models can generate instances of what they have learned based on a noise input vector that can be interpreted as a latent representation of the data at hand. By manipulating this latent representation and examining its impact on the output of the generative model, it is possible to draw insights and discover specific patterns related to the class to be predicted.'' We postulate that these insights and patterns from GANs can be integrated with XAI to provide richer explanations, yielding more knowledge about the classes and data in question.

E. RISKS OF UNEXPLAINABLE AI IN CPS

Many risks arise when AI is implemented in CPS without explainability. Unintended consequences of unexplained AI in CPS include undetected bias, improper use of the system due to lack of trust, unfair results, failure to correct and update the AI-based system as flaws are discovered, security risks, lack of accountability, and lack of appropriate human oversight. Much of the recent interest in interpretability and XAI is driven by the need for accountability and transparency of AI-based decisions as well as the risks of cybersecurity attacks against AI-based cyber-physical manufacturing systems [50]. Explainable AI also helps to ''minimize the cost and consequences of poor decisions'' [50]. In CPS, poor decisions can create financial loss, loss or waste of materials, products of unsatisfactory quality, failure to meet requirements, system shutdown, and accidents.

At industrial CPS manufacturing locations, workers need to be protected from accidents; the company and management must also be prepared for legal consequences of such accidents if they occur [51]. Accidents can occur within the system, in the surrounding physical environment, or as a result of a decision the system has made. For example, Li et al. [48] discuss a possible case in which a poisoning attack occurs in an industrial manufacturing cyber-physical system. The attacker explores the system, creates specially tailored data, feeds the poisoned data to the system, manipulates the communication process, and uploads illegal parameter values. When this attack occurs, any abnormality may cause the entire production line to fail [48]. In [49], an example is given in which an adversarial example attack is used to train systems' DL-based CPS to incorrectly ''identify the type of electronic devices (such as capacitor, inductor, and transistor) on circuit boards. A mistake that identifies a capacitor as an inductor may cause important chips to burn out.'' This could lead to equipment failure, a dangerous prospect in safety-critical applications such as motor vehicle circuitry or healthcare CPS.

Adding explainability to CPS paves the way for better cybersecurity. It offers users the opportunity to monitor systems and manually examine what is happening if a suspected attack has occurred. Without explainability, it would be difficult for a human user to evaluate for such attacks in an AI-based CPS. With XAI, the user can run inquiries, assess the feature importance, and evaluate the reasoning behind decisions. These steps may be used to augment cybersecurity measures by learning more about adversarial attacks and finding ways to be prepared for future attacks. New XAI methods can be developed specifically to search for and explain poisoning or other types of attacks.

Wickramasinghe et al. [1] state, ''Many modern critical infrastructures have CPSs at their core. Therefore, these systems are highly vulnerable to various attack vectors. Consequently, maintaining the safety and security of CPSs is a primary focus.'' The authors outline an XAI-integrated approach in which ML-based anomaly detection systems are developed to analyze system behavior and create clusters which show likely normal behavior. Interpretations are developed for each cluster. The explanations can then be analyzed by domain experts, who can determine whether the clusters represent normal behavior, find the most important features correlated with normal behavior in the system and in each cluster, and perform other analyses. These steps combined with explanations enable the domain experts to enhance system security, reduce data dimensionality, identify important data features, and avoid bias [1]. XAI is what makes all of these steps possible. Table 3 provides a side-by-side comparison of AI, XAI, and CPS together, discussing the goals, definitions, and actions of each in context of the relationships between the three. Fig. 3 provides a visual representation of this information.

Studies have demonstrated the successful implementation of XAI for attack detection in cyber-physical systems [8], [9], [41], [50]. XAI also drives cybersecurity improvement in CPS by assuring accountability. If a breach or attack occurs, XAI associates ''each activity with a specific individual rather than a group or an ID. This compels security frameworks to become more resistant to cyber attacks'' [54].

IV. EXPLAINABLE ARTIFICIAL INTELLIGENCE IN CYBER-PHYSICAL SYSTEMS

A. BARRIERS TO ADOPTION OF AI IN CYBER-PHYSICAL SYSTEMS

In spite of the benefits of implementing AI and CPS, many stakeholders are reluctant to adopt these technologies, especially in highly regulated or high-stakes sectors [16], [28], [29]. In order for stakeholders to be willing to implement them, the systems must be trustworthy and explainable: ''Opaqueness affects trust in the system, a factor that is critical in the context of decision-making'' [17]. Trust can be defined as follows: ''An entity's decision to interact with others is an act of trust…the trustor relies on and places its trust in the trustee to accomplish the task as agreed upon,'' given that a trustor is the subject which places its trust in a target entity called a trustee [55]. Trust facilitates interaction with other entities in uncertain environments and helps the

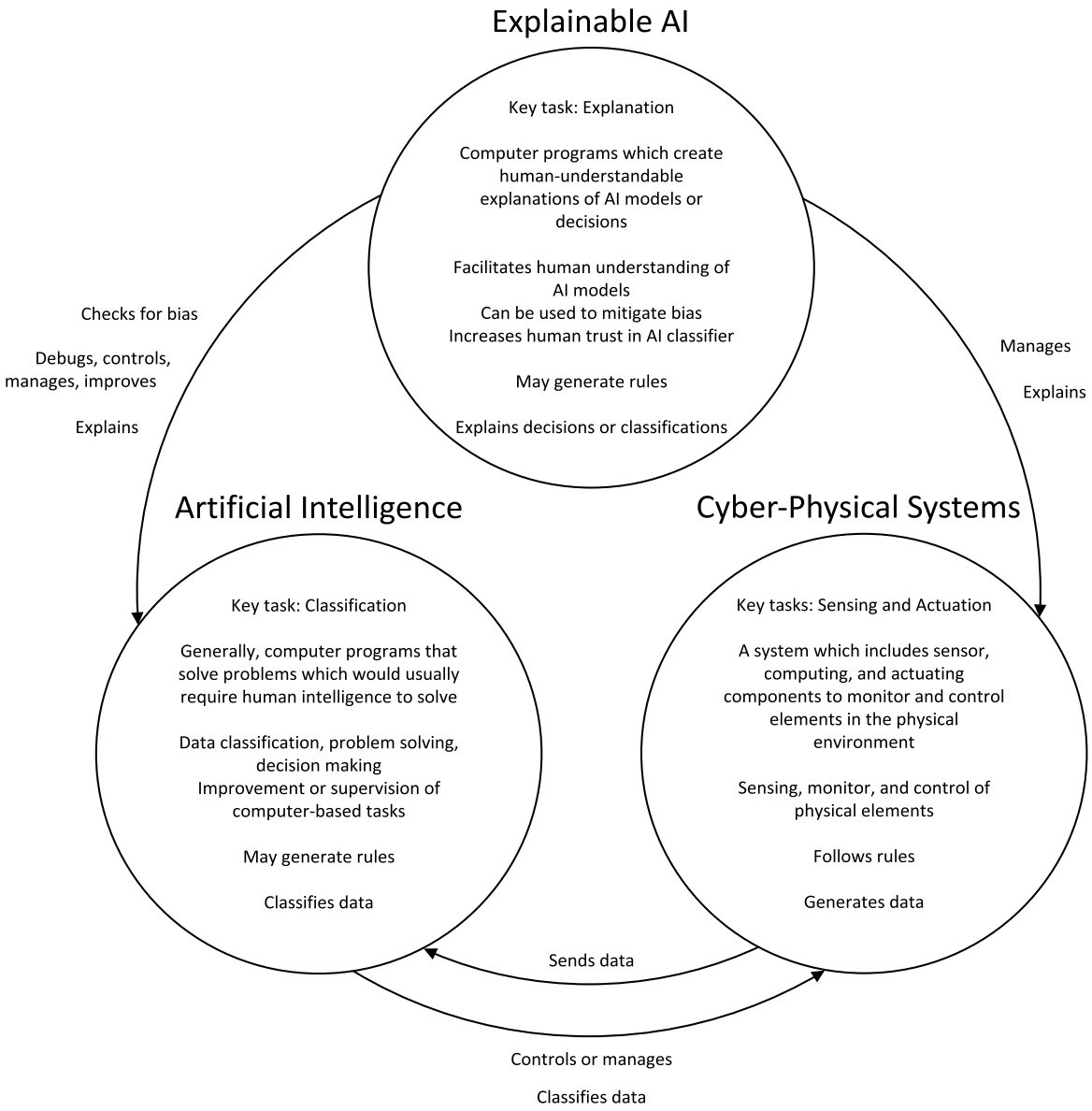

FIGURE 3. A comparison diagram showing the characteristics and connections between AI for classification, XAI, and cyber-physical systems.

trustor to predict an entity's behavior. Trust has a powerful effect on the extent to which human operators are willing to rely on automation [56].

In some cases, human trust and reliance on AI-integrated systems will depend on how the system performs in sensitive situations [27]. For example, cyber-physical AI-based battle systems will be required to make many sensitive decisions. It is therefore important not only to make AI systems that are as reliable and accurate as possible, but also to make them as trustworthy as possible so that the humans in the loop can have the benefit of using AI to make the best possible decisions. AI can integrate information from many different information sources and types and in much higher quantities than a human could possibly analyze [57], [58]. In battle or rescue situations, for example, time is of the essence. A human user must be able to take advantage of the helpful information available from the AI model. Understanding how the system works can help to increase the user's trust in it [12].

According to Mendia et al. [40], 90% of AI models never reach production in real-world industrial solutions. This is due in part to complexity and performance, but more importantly, it is due to explainability concerns [40]. The AI-derived recommendations need to be explained by the system as well as understood and trusted by the users. Ferraro et al. [28] aver that humans "are reluctant to adopt techniques that are not directly interpretable, tractable, and reliable," especially as demands for ethical AI continue to increase. For example, in the medical field, practitioners require a great deal more information from an ML-made diagnosis prediction than just a binary value. Understanding

TABLE 3. Comparison of AI, XAI, and CPS.

	AI for Classification	XAI	CPS
Definition	Can generally be defined as computer programs that solve problems which would usually require human intelligence to solve [52]. Commonly used for decision making or data classification.	Defined as methods used to create human-understandable explanations of the logic behind decisions made by AI models [1], [53]	Defined as a system which includes sensor, computing, and actuating components to provide monitoring and control of physical elements [2]–[4]
Goal	Generally, to classify data or make decisions, predictions, or recommendations based on input. Can be used for improvement or supervision of computer-based tasks [53]	To offer reasoning for decisions made by AI classifiers, provide explanations and facilitate human understanding of how the classifier works, mitigate bias, and increase human trust in the AI classifier [1], [53]	To sense what is going on in the environment, provide control, and affect elements in the environment [2], [4]
Actions Performed	Classifies data into categories; makes decisions, predictions, or recommendations	Creates explanations for AI-based decisions or classifications	Senses elements in the physical environment and makes changes (performs actuation)
Relationship to Rules	May generate rules	May generate rules	Follows rules
Connected Relationships between AI, XAI, and CPS	Can be used in the control unit of a CPS to determine what actions to perform, classify measurements, etc.	Can be used to manage and explain AI. Can be used to debug, control, explain, learn more about, or improve CPS	Can be managed by AI or XAI; can be explained or improved by XAI. Sensors can input information into AI or XAI models for decision making, control, or classification.
Relationship to Data and Information	Classifies data	Explains decisions or classifications	Generates data

the reasoning behind AI-based predictions is crucial for patient safety. Cartolano et al. [21] emphasize this, stating, "there are some fields of application in which [it] is extremely dangerous to be confident in predictions without an explanation, for instance in medical science where it has been proven that not interpretable models could potentially cost human lives."

XAI can be used to improve existing cyber-physical systems. As the field of explainable artificial intelligence grows, making more models available, stakeholders will be more willing to implement it in their systems and thus gain any resulting improvements from XAI such as decreased production time, decreased system downtime, and increased profits [13]. The implementation of AI in industry can have great benefits such as allowing systems to make decisions and take action quickly, often acting automatically without the need for direct human control [59]. Such implementation requires AI to not only work as intended but also to be safe, trusted, ethical, and legal. To achieve this responsible design, AI must be explainable and accountable [59]. Stakeholders must be able to justify AI-recommended decisions; XAI provides a basis for this.

Some AI methods currently utilized in manufacturing and industry, typically ML- or DL-based, are used to reduce downtime by predicting required maintenance [13]. In systems engineering, AI can be used for modeling, requirements modeling, monitoring coding, and testing [59]. Adding XAI to any of these applications can make them easier to understand, making it more appealing for stakeholders and users to implement AI.

B. AI, XAI, AND CPS IN INDUSTRY 4.0

AI and CPS together are forging the way into Industry 4.0. Ahmed et al. [13] name AI as the leading component transforming current industry into the Fourth Industrial Revolution; it allows for self-monitoring, diagnosis, analysis, interpretation, and execution of many other autonomous tasks in cyber-physical systems. Predictive maintenance is another key element of Industry 4.0, aiming to reduce downtime and maintenance costs. By correctly predicting and maximizing equipment's remaining useful life (RUL), XAI can be harnessed to optimize maintenance costs, saving up to 60% of these expenses through correcting faults [28]. Existing DL models used for predictive maintenance are usually black box models. XAI can be used to comply with regulations in highly regulated or high-stakes fields, increase user trust, and allow dynamic system improvement. XAI can also show the strengths and weaknesses of the model's decision making process in a clear, human-understandable way [28]. An example of explainable AI implemented in cyber-physical manufacturing systems is provided in Fig. 4. This use case diagram shows ways in which different users, including an XAI system operator, supply chain manager, internal auditor, programmer, expert user, and repair technician, may interact with an XAI model integrated with its cyber-physical manufacturing system. The XAI system can be used to understand decisions, check and update rules, verify compliance, ensure that necessary supplies are available, discover predictive features, acquire transferable knowledge, and perform many other tasks.

According to Sofianidis et al. [17], "AI is currently the most disruptive digital enabler of the Industry 4.0 era". Deep learning models are extensively used in existing industries, achieving excellent results in fault diagnosis, fault classification, and prediction of key performance indicators [28]. With CPS and backbone technologies such as AI, industrial IoT, and cloud computing, Industry 4.0 is supporting the development of mass customization, flexible production lines, predictive maintenance, digital twins, and Zero Defect Manufacturing [17]. AI and CPS together can allow systems

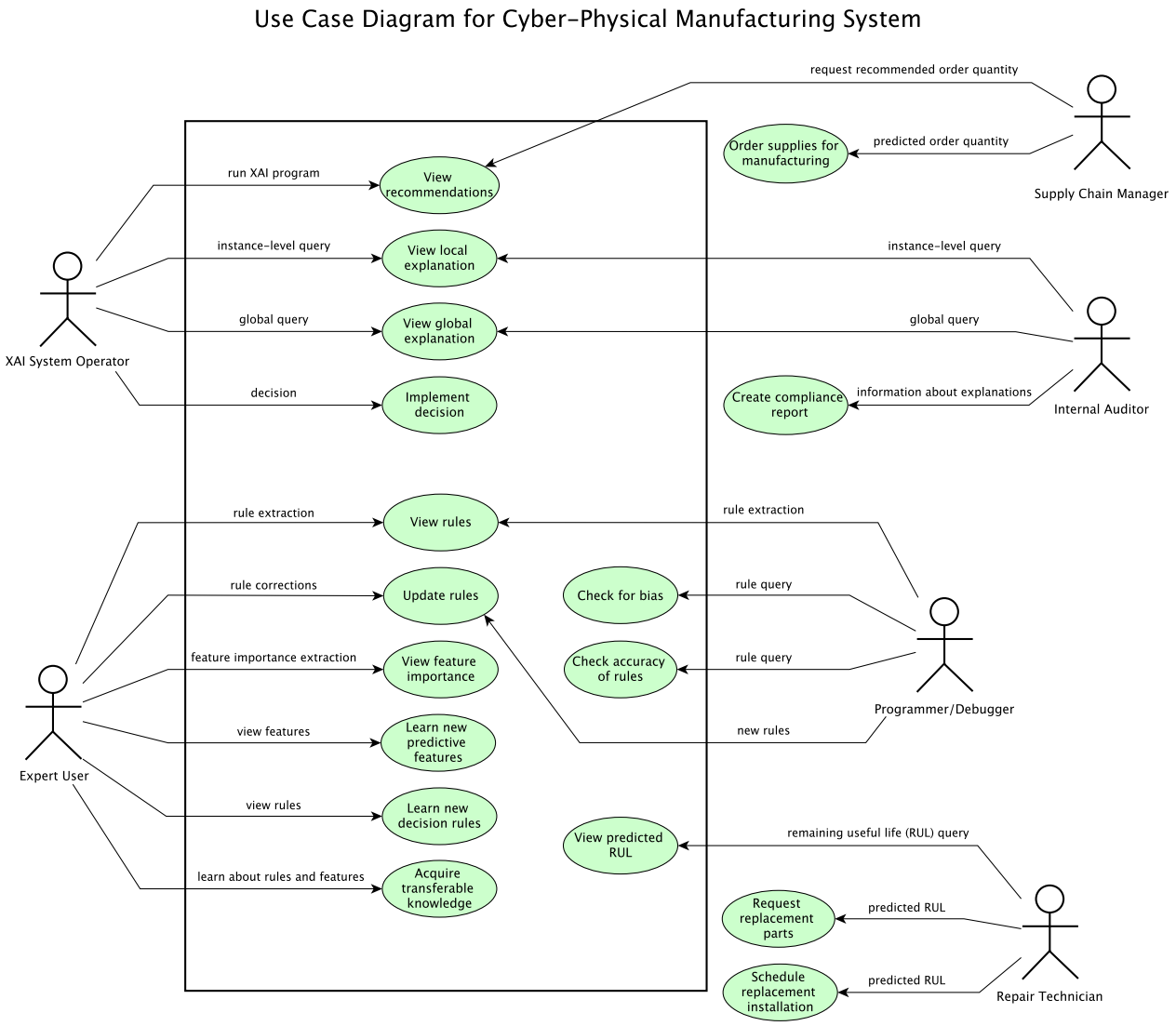

FIGURE 4. **Use case diagram demonstrating implementation of XAI for a cyber-physical manufacturing system.**

to adapt rapidly to global changes in supply and demand, sociopolitical situations, environment, and finances [5]. Even while adapting, these systems can maintain optimization and process control.

XAI methods such as SHAP and LIME can be used to predict the remaining useful life of machines in cyber-physical systems, thus preventing waste of systems that are still functional and decreasing downtimes by having replacements ready when they are needed [60]. XAI for RUL prediction can also prevent system unavailability, downtime, and data loss due to failed hard drives in a CPS [28]. XAI can be utilized to improve user learning performance in heavily regulated applications such as autonomous vehicle guidance or military applications [28].

As time passes, more cyber-physical systems are integrated with AI-based applications [5]. Artificial intelligence not only enables the creation, running, and optimization of cyber-physical systems; it also allows multiple cyber-physical systems to connect, collaborate, analyze, and provide system cognition. As technology advances, cyber-physical systems can expand into "distributed, large-scale, cooperative, and flexible automation systems" [5]. The global paradigm changes in industry and computing necessitate the integration and synchronization of multiple cyber-physical systems.

For human-in-the-loop cyber-physical systems, it is of particular importance for the human to understand the reasoning of the artificial intelligence so that they can use the information it offers [1]. Explainable artificial intelligence creates a way for the human in the loop to understand and trust the decision-making process of the AI model.

The explainability problems with AI in CPS typically result from the use of black-box models, such as DL or artificial neural network (ANN) models [61]. Weyns et al. [29] remark that one of the key challenges in cyber-physical systems is "to provide methods and techniques for understanding data driven models to support data driven applications and

to transfer knowledge to other settings," and XAI can be used to meet this challenge. Interactive visualization can also improve the explainability and interpretability of XAI models.

C. CYBERSECURITY, INTRUSION DETECTION SYSTEMS, AND XAI IN CYBER-PHYSICAL SYSTEMS

The ever-growing number of kinds of cyberattacks makes it critical for intrusion detection systems to be extendable and adaptable [14]. Capuano et al. [32] state, "According to AV-Test Institute, more than 1 billion malware programs are out there, and 560,000 new pieces of malware are detected every day." Attacks such as Trojans, worms [2], data poisoning, model poisoning, backdoor [48], Mirai botnet, Denial of Service (DoS), Distributed Denial of Service (DDoS), scanning [62], information stealer, rootkit, downloaders, ransomware [32], and many others flood networks, putting network-connected cyber-physical systems at risk.

Most of the machine learning techniques used in intrusion detection systems are not explainable. Understanding the logic behind decisions and recommendations from AI, ML, and DL models is of great importance; governments are even beginning to implement legislation to require explainability as a part of AI-based applications. Marino et al. [14] discuss the strong need for transparency in IDS, especially for the engineers developing them. This explainability facilitates knowledge and data discovery, debugging, diagnosis, and trust in the model.

According to Capuano et al. [32], most of the analyzed explainable intrusion detection systems use post-hoc explanations, especially SHAP. Most of the inherently interpretable methods used decision trees. The researchers remark that future research should explore the countering of adversarial attacks in which human-machine collaboration is necessary; explainability is fundamental to prevent these kinds of attacks [32]. Houda et al. [62] use RuleFit, LIME, and SHAP with deep learning models to explore local and global explanations for interpretation of decisions made by the DL models. They classify the NSL-KDD and UNSW-NB15 datasets as Attack or Normal; these datasets include Root to Local, User to Root, Probe attacks, DoS, DDoS, exploits, worms, reconnaissance, fuzzers, shellcode, analysis, backdoors, and generic attacks.

Amarasinghe [63] performed an experiment in which an ANN based anomaly detection system called NN-ADS was used to classify intrusions as either normal or a DDoS attack. Explanations were created as linguistic/numeric summaries for all 78 available features in the tested sample of the CICDS2017 dataset. The experiment was able to classify the samples with 98.71% testing accuracy. The author then performed validation using adversarial examples. The author also implemented Deep Self-Organizing Maps (DSOM) for learning the behavioral patterns of cyber-physical systems [63]. The researcher used DSOM on the KDD-NSL dataset for prediction of the categories of Normal, DoS attacks, and Probe attacks, providing linguistic and visual explanations from the output layer self-organizing map to find the learned behavior patterns.

D. RELATED RESEARCH

This section provides a discussion on recent research related to XAI and cyber-physical systems. We review studies that evaluate applications to smart cities, non-terrestrial networks in smart healthcare, smart homes, the use of XAI to create powerful intrusion detection systems for CPS, and the importance of information extraction from explainable AI-based IDS to create stronger cybersecurity systems. We then discuss studies about XAI in manufacturing, industrial CPS, and systems engineering. Lastly, we consider research which shows the importance of learning and flexibility in explainable AI for CPS. Table 4 provides a summary of relevant literature including key contributions and limitations.

A survey by Javed et al. [34] provides an in-depth view of the current and future needs, driving forces, technical challenges, recent findings, and implementations for XAI in smart cities; the authors state that their review is "a first-of-its-kind, rigorous, and detailed study to assist future researchers in implementing XAI-driven systems." Its comprehensive review of recent and future developments maintains a focus on conceptual applications including "XAI technology use cases, challenges, applications, and possible alternative solutions" [34]. The researchers also highlight the following critical issues for XAI in smart cities: integrity, compliance, monitoring and audit, ethical and privacy issues, availability, authentication and authorization, standard specifications, scalability, and justice and reasonableness of AI-based decisions.

Pattepu et al. [64] contribute a new XAI-based cooperative relaying scheme for non-terrestrial network (NTN) CPS called Incremental Multi-Antenna Relays Cooperation with a Hybrid relaying scheme (IMARC-H). The research is focused on applications in smart healthcare networks. Their method outperforms other tested relay methods such as IMARC-AF, IMARC-DF, IMARC-r1DF-r2AF, and IMARC-r1AF-r2DF in terms of minimizing the bit error rate (BER) and probability of outage [64].

Houze et al. [65] propose D-CAN, a decentralized algorithm and generic decentralized architecture which "enables XAI functions to be extended and updated when devices join and leave the managed system dynamically." The authors demonstrate the effectiveness of their method through smart home case studies. D-CAN is intended to create user understandable explanations using logical steps for unexpected situations in the smart home; it also provides a stepwise guide to resolve the situation [65]. The method is designed to be generalizable to other decentralized CPS with similar challenges including openness and high dynamism. Critically, this method is designed to "ensure [a] seamless integration of components at runtime and adapt to larger systems" [65]. This research is important in that it addresses

the large, decentralized, flexible nature of many cyber-physical systems. CPS may have parts added, removed, updated, or upgraded at any time; XAI solutions must be able to handle these changes as they occur to prevent any security vulnerabilities.

Amarasinghe [63] developed new explainable ANN-based anomaly detection systems for CPS. Furthermore, these explainable models are intended to increase the trust of human operators in these ANN-based systems. The author defines the set of basic requirements for explainable anomaly detection systems for CPS, contributes a methodology for creating summaries of the knowledge gained by the supervised CPS anomaly detection system, adds a methodology for validating these summaries, offers an unsupervised NN method for learning CPS behavior, and provides a method to explain this behavior both visually and through language. This research ''serves as a framework that can be expanded to develop trustworthy ANN-based [Cyber-Physical Anomaly Detection Systems]'' [63].

Cyberattacks on network-connected computing devices pose risks to information security, system performance, and privacy among many others [41]. However, the risks from intrusions in cyber-physical systems extend beyond the network and data into the physical realm. Sensors and actuators in CPS can malfunction; physical objects can be damaged or broken, and even physical safety can be put at risk [41]. Although machine learning has been implemented for intrusion detection in CPS, the lack of labeled data on new attacks creates problems in systems' ability to detect these intrusions. The authors posit that deep learning can be combined with XAI to create sophisticated new IDS with minimal complexity [41]. When IDS for CPS fail, this can cause system failure, loss of equipment, loss of products, and loss of sensitive data such as proprietary information and trade secrets [41]. One of the basic problems of creating cybersecurity for CPS is taking into account the physical processes of the system; these processes can be altered or made to operate in an unsafe manner by cyberattacks. As CPS become larger, grow more complicated, and gain more complex and numerous interconnections, additional attack vectors are created.

Alqaralleh et al. [41] created and tested a new intrusion detection system for CPS called XAIIDS-FSDVAE, short for XAI-based Intrusion Detection System using Feature Selection with Dirichlet Variational AutoEncoder. Their approach combines many subprocesses for optimal results, including preprocessing, Coyote Optimization Algorithm (COA) based feature selection and reduction of computational complexity, Dirichlet Variational AutoEncoder (DVAE) based classifiers, and parameter optimization with Manta Ray Foraging Optimization (MRFO) [41]. Tested on the CICIDS-2017 dataset, this method outperformed Dirichlet Variational AutoEncoder (DVAE), Parameter-Tuning Deep-Stacked AutoEncoder (PT-DSAE), Decision Tree (DT), Random Forest (RF), Fog-based Intrusion Detection with Generative Adversarial Network (FID-GAN), Multivariate Anomaly Detection with Generative Adversarial Network (MAD-GAN), and Adversarially Learned Anomaly Detection (ALAD) in terms of accuracy, precision, recall, and F-measure, all of which reached above 99% using the authors' new method.

The burgeoning field of XAI has enabled more robust intrusion detection methods to be developed for CPS. Almuqren et al. [9] developed an XAI method for intrusion detection in CPS and tested it on the NSLKDD 2015 dataset. It outperformed all other methods tested (FURIA, AE-RF, Forest-PA, WISARD GSAE, and LIB-SVM) in accuracy, precision, recall, and F1 score. These results show that XAI methods can provide superior results in intrusion detection while still maintaining explainability.

Marino et al. [14] contributed a new approach which generates explanations for wrongly classified instances in intrusion detection systems using adversarial machine learning. This is done by using an adversarial approach ''to find the minimum modifications (of the input features) required to correctly classify a given set of misclassified samples.'' Then, the magnitude of the changes required in each feature to yield the correct answer is displayed visually, thus showing the most relevant features behind the misclassification [14].

Mongelli [66] found that information extraction from artificial intelligence may be more important in maintaining cybersecurity than the AI model's prediction ability. XAI can be used for rule extraction and explanation to provide insight into the cybersecurity problems at hand. For example, important features, values, and rules can be rated, extracted, implemented in program or system design, and then tested to ensure safety or cybersecurity.

Sofianidis et al. [17] provide an excellent, in-depth review of XAI in manufacturing. They note that AI implementation in manufacturing applications can increase efficiency of production, performance, and safety. However, without explainability in these AI-based systems, trust is compromised; trust is critical in decision-making. This article reviews XAI methods and techniques, taxonomies, measures of evaluation, technique classifications, applications, use cases, and current challenges [17].

Mendia et al. [40] contributed an XAI model called NAIA, ''a novel tool designed to characterize, in a non-supervised, human-understandable fashion, the nominal performance of a factory in terms of production and energy consumption.'' This tool can be integrated with cyber-physical production systems to trace and analyze energy consumption, facilitating anomaly detection and inefficiency detection and providing human-understandable information about the root cause of the problem. This study stresses the importance of taking into account the domain knowledge and target audience to design AI-based industrial solutions which can be effectively implemented in the real world [40].

XAI is a key part of the systems engineering process [13]. Systems are rapidly being developed, advanced, and altered as Industry 4.0 continues. XAI is a vital tool for gathering required information and adapting existing systems, enabling

TABLE 4. Related research.

Ref.	Topic	Key Contributions	Limitations
[66]	Smart transportation cybersecurity, safety, and cyber-resilience in vehicle platoons	XAI method to detect and apply countermeasures to packet falsification attacks in vehicle platoons	Needs additional real-world testing to verify method
[13]	XAI, Industry 4.0, cyber-physical systems, IoT, cloud computing, technologies in industry	A survey of industrial technology, AI, XAI, cyber-physical systems, applications in Industry 4.0, and related challenges	Does not demonstrate methods of maintaining security and cyber-resilience
[7]	XAI methods used with medical CPS	A review of existing XAI methods as applied to EEG, EMG, and ECG	Does not include other types of medical CPS beyond EEG, EMG, and ECG
[9]	CPS security and intrusion detection	High-accuracy intrusion detection and classification for CPS; uses LIME as XAI; provides extensive evaluation metrics	Implementation is on simulated/experimental network; does not include real-world network dataset implementation; does not include clustering or outlier removal
[41]	XAI for intrusion detection in cyber-physical systems	Developed a high-performing XAI-based intrusion detection system for use in CPS; includes evaluation metrics and comparison with other methods	Does not include outlier detection or cluster-based approaches
[63]	Anomaly detection for CPS using explainable ANNs	Develops supervised and unsupervised ANN-based anomaly detection systems for cyber-physical systems; develops methods to quantitatively validate the derived explanations	Does not provide contrastive explanations or individual feature-based clustering explanations
[28]	Predictive maintenance, XAI, hard disk drives, LIME, SHAP	XAI for RNN and LSTM-based prediction of remaining useful life of hard drives; spatial and temporal features; displays generated dashboards	Does not include validation of the explanation framework with human users
[65]	CPS with decentralized decision making, smart home systems	A new decentralized XAI method for CPS; implemented on a smart home CPS simulator	Does not have objective measures for explanations given in proof-of-concept simulation
[34]	XAI for smart cities	Survey of current XAI methods, use cases, applications, and future research areas for smart cities	Does not discuss XAI as combined with recent security frameworks in the application of smart cities
[14]	XAI in intrusion detection systems, adversarial machine learning, cybersecurity in local area network environments	Performs input feature modification to correctly classify previously misclassified samples; provides explanatory visual output	Uses simple ML models including linear classifier and multilayer perceptron classifier; uses a non-recent dataset (NSL-KDD99); further validation could be performed using more complex machine learning models
[40]	Industrial CPS, anomaly detection, energy consumption	Demonstrates a new tool to clearly explain the production and energy consumption of a factory, facilitating anomaly and inefficiency detection; includes real-world industrial factory implementations	Does not include detection or adaptation of data shift or concept drift; model requires a priori knowledge about system hierarchy and factory processes
[64]	Healthcare systems, network relaying schemes	XAI implementation for optimal communication in healthcare system non-terrestrial networks using XAI-based cooperative relaying scheme	Does not include dynamic signal-to-noise ratio; current implementation has a limited number of cooperative nodes
[15]	Robot navigation in dynamic environments, XAI	Sensor-based robot navigation; use of decision trees, expert policies, and policy extraction to modify existing policies; modification of learning algorithm without retraining	Limited problem-solving applications are tested (blocking, oscillation, freezing); obstacles cannot be moving faster than the maximum speed of the robot
[17]	XAI in manufacturing, review of XAI	Provides a thorough review of XAI and desiderata, specific evaluation measures, list and classification of XAI techniques, manufacturing use cases, and applications in manufacturing	Does not discuss cyber-resilience
[1]	Unsupervised learning, XAI methods, self-organizing maps (SOM)	Provides an SOM-based XAI model which uses clustering methodology and offers both local and global explanations; explainable AI for unsupervised machine learning models; review of existing XAI methods for use with unsupervised ML	Results may change when different distance measures are used

flexibility in an environment of rapid change. Together, "AI and XAI enable automatic and real-time implementation of these intelligent systems and applications" [13]. AI empowers autonomous machines and task completion while XAI adds human-understandable explanations. Existing cyber-physical systems in manufacturing and industrial systems typically produce large amounts of data which are not properly captured and analyzed. Improved big data analytics will enable XAI implementations which better facilitate the meeting of Industry 4.0 goals [13].

Ferraro et al. [28] implemented an experimental model using LSTM on the Backblaze dataset to predict the remaining useful life (RUL) of hard drives; the authors then implemented SHAP and LIME to provide explanation and evaluation of the RUL predictions of the model [28]. They found that SHAP outperformed LIME in most of their evaluation metrics, although LIME provides an explanation tool that shows the "contribution of each feature in all time instants within a time window" [28]. The model utilizes three-dimensional input data including samples, timesteps, and features.

Roth et al. [15] developed "a novel sensor-based learning navigation algorithm to compute a collision-free trajectory for a robot in dense and dynamic environments with moving obstacles or targets." This method relies on an expert policy based on deep reinforcement learning; it is also integrated

with a policy extraction technique which generates the learned policies as a decision tree format. The decision trees can be used to understand, analyze, and change policies as needed to improve performance [15]. The rule extraction and analysis can be used to solve problems, for example, to decrease immobilization frequency, improve smoothness, manage cases which lead to failure, or increase reliability, all without the need to retrain the algorithm.

One of the most important reasons for AI to be combined with CPS is that it allows cyber-physical systems to learn. ''The use of learning capabilities allows CPSs to interact and analyse their environment, learn from patterns, and perform highly complex prediction tasks'' [7]. Not only can CPS do what they are intended to, but with AI they can also learn to do it better: for example, more safely, more quickly, with fewer resources, or with certain desired parameters or conditions. However, these abilities drive the need for transparency and understandability within AI-based CPS, especially in critical fields such as medicine. Indeed, Alimonda et al. [7] state that understandability and transparency are of equal importance to performance in medicine, enabling users to trust, comprehend, and understand errors in systems. In their review, the authors discuss many studied applications of XAI in medical cyber-physical systems such as classification of heartbeats, heart disease, electrocardiogram (ECG) signals, or arrhythmia; detection of finger movements, stress, or seizures; and emotion recognition, with many more applications to come [7]. The authors name XAI as the ''new frontier to obtain more reliable CPS which use black-box learning models.''

Wickramasinghe et al. [1] propose a new model-specific XAI method: ''a Self-Organizing Maps based explainable clustering methodology which generates global and local explanations'' which identifies the most important features in the model's decision making. The researchers demonstrate the capability and strengths of their model through comparison with other unsupervised XAI methods in terms of qualities, capabilities, applications, computational complexity, and limitations. The authors state that tremendous amounts of unlabeled data are generated every day by cyber-physical systems; relying on supervised data alone is not enough to keep up with the new and changing data outputs of CPS [1]. Unsupervised AI methods can be implemented to take full advantage of the available data and make data-driven decisions.

E. CHALLENGES AND RECOMMENDATIONS FOR CURRENT XAI FOR CYBER-PHYSICAL SYSTEMS

Cyber-physical systems are being implemented to automate complicated physical processes. This increases the complexity of the systems, which necessitates explainability to understand the function and behavior of the system [6]. However, most of the commonly used XAI methods for explaining CPS behavior ''usually overlook the impact of physical and virtual context when explaining the outputs of decision-making software models, which are essential factors in explaining CPS' behavior to stakeholders'' [6].

Jha [6] reviews existing XAI methods for explaining CPS behavior, discussing shortcomings and research directions. Current methods are not context-aware, need intelligible visualization tools, and need actionable explanations from visualizations. The XAI techniques used in CPS still need further research to become understandable for many types of users, not just AI experts. The author recommends improving current explanations by providing contextual information using semantic technologies and adding ways to enter and incorporate user feedback [6]. In addition, the author states that the explanation delivery mechanism can be improved by including counterfactual explanation methods and knowledge graphs. This will provide explanations that are easier to understand.

Most of the research in XAI has been focused on supervised machine learning methods [1]. However, cyber-physical systems generate large amounts of unlabeled data at a quick pace. Manual labeling is costly, time-consuming, and requires domain expertise [67]. Unsupervised machine learning allows the existing data to be analyzed without manual labeling and can also help to avoid bias [1]. By classifying data without prior existing labels, unsupervised learning uses a more neutral method of classification than data labeled by humans or datasets automatically labeled based on biased samples.

F. XAI AND CPS IN SYSTEMS ENGINEERING

Beyond applications with existing cyber-physical systems, XAI can be used in the system development lifecycle [59]. It can be incorporated into each phase at design-time and runtime, providing explanations and accountability at every step. The use of AI in mission-critical operations raises questions about verification and validation of any AI-generated recommendations [59]. XAI makes it possible for users and stakeholders to provide and check justifications for those decisions, assisting in the verification and validation processes.

In some high-stakes applications or highly-regulated environments, there must always be a human present to review and verify any decisions made by the AI model [59]. XAI can make this task easier by automatically providing a rationale as to why the recommended decision is likely the correct one. However, human users must be careful to not trust the XAI model too much to the point of ignoring their own judgment or assuming that because the generated explanation is plausible, it must be true. It is possible for XAI models to produce explanations that make sense but are in fact false [27].

G. THE DANGERS OF BIAS IN CPS

Biased training data, or training data that otherwise does not accurately represent the real world, can have dangerous results for cyber-physical applications as well. For example,

training a self-driving car only on a test track will leave it dangerously unprepared for real-world driving scenarios [27]. If AI-based cyber-physical systems are not well prepared for different types of adversarial attacks, they may behave in unexpected, undesired, or dangerous ways. Therefore, stakeholders need to know not only how the AI-integrated CPS works and how to make it work well, but also how it can go wrong [27]. XAI not only teaches users to understand how AI networks function, but it can also allow developers to find errors, flaws, biases, and opportunities for improvement in existing AI designs. Based on what developers learn from XAI, they can improve their models by fine-tuning hyperparameters, altering training and testing datasets, performing feature engineering, and modifying architecture [12]. This will improve the safety and functioning of cyber-physical systems.

H. XAI IN MEDICAL CPS

In the medical field, cyber-physical systems are of ever-growing importance. According to Alimonda et al. [7], transparency and interpretability are equally important as the performance of the model. These characteristics foster trust by allowing users to understand the behavior and possible mistakes of the system. Thus, XAI is a necessary component of medical CPS to create transparency, interpretability, trust, and appropriate decision making.

In medical CPS, XAI can "help clinicians by providing information about features that contribute to the outcomes and their importance in the decision making process, in order to have a better understanding of how the system works" [7]. The goal of XAI in medical CPS is to make clear, understandable output that increases the confidence of users as they make decisions. Some medical systems which benefit from integration with XAI-based cyber-physical systems include electroencephalogram (EEG), electrocardiogram (ECG), and electromyography (EMG). Machine learning can be used with these systems to find patterns associated with particular diseases and make predictions about new inputs [7]. Another type of cyber-physical system used with DL is pathological imaging [34]. In such cases, XAI visualization tools can be used to produce a map overlay that emphasizes the parts of a visual input which were most important in determining the output [7]. XAI visualization methods are often used for images or spatial data from EEG, EMG, ECG, or electrooculogram (EOG). Common XAI methods used with EEG in biomedical systems include SHAP, NBET, LIME, feature importance, and GRAD-CAM. GRAD-CAM, LRP, STF, and SHAP have been used for EMG; GRAD-CAM, feature importance, SHAP, and LRP have been applied for EEG [7].

Medical CPS will increasingly involve multiple sensors not just in the clinical environment but also at home [68]. For example, a patient with a spinal injury may carry a portable sensor to measure their movement as they climb stairs. Other sensors may monitor posture, activity level, sleep, blood pressure, pulse rate, etc. Sonntag et al. [68] developed a system in which active and passive user input data are gathered and analyzed, knowledge integration is performed, and machine learning-based decision support is offered. The addition of XAI to such a system would allow for improved regulatory compliance, ethical decision making, and understanding of the rationale behind treatment recommendations.

Computerized clinical decision support systems (CDSS) gather information and assist healthcare professionals to make good decisions, supporting the process of providing healthcare and improving patient health outcomes [34]. These systems can provide assistance with diagnosis, individualized treatment recommendations, assistance in hypothesis development for challenging cases, prediction of the response to treatment, prognosis, and risk-based prioritization [33], [34]. These systems offer more options to the provider and patient while reducing mistakes.

When AI is added to CDSS, it runs the risk of including bias, making inaccurate predictions, or providing skewed results if the training data are biased [34]. XAI is needed in CDSS to enable better monitoring for bias, understanding of the logic behind recommendations, and management of the system. Understanding of this logic is especially critical when recommendations could put people in danger [34].

I. THE IMPORTANCE OF UNDERSTANDABILITY IN CPS

The critical nature of understandable explanations in smart systems can be illustrated with an example in smart agriculture. In a study by Cartolano et al. [21], XAI was used to predict which crops are suited to a location given its soil and environmental conditions. The researchers demonstrated that by using XAI charts, farmers and agronomists who are not ML experts can understand the reasoning behind the model's prediction of a certain crop for a given location or data instance. The users can also learn which are the most important feature combinations that led to a given type of prediction. This is especially important when initially implementing models in the real world, where the expected performance achieved in a controlled, artificial setting may not represent the real-world performance. The farmers' and agronomists' expert knowledge can be compared with the results of the XAI predictions, allowing the users to learn about the model, learn more about their domain of expertise, and to check for bias or errors in the model [21]. This could prevent harmful economic consequences of crop loss due to incorrect predictions made by an AI model.

J. REQUIREMENTS ANALYSIS FOR XAI-INTEGRATED CPS

Requirements analysis is a foundational process in building XAI-based cyber-physical systems. It is used to list and define the necessary elements and resources of an effective system, ensuring successful design, development, and deployment [69]. Leveraging this process ensures that the system meets stakeholder needs, fulfills user expectations,

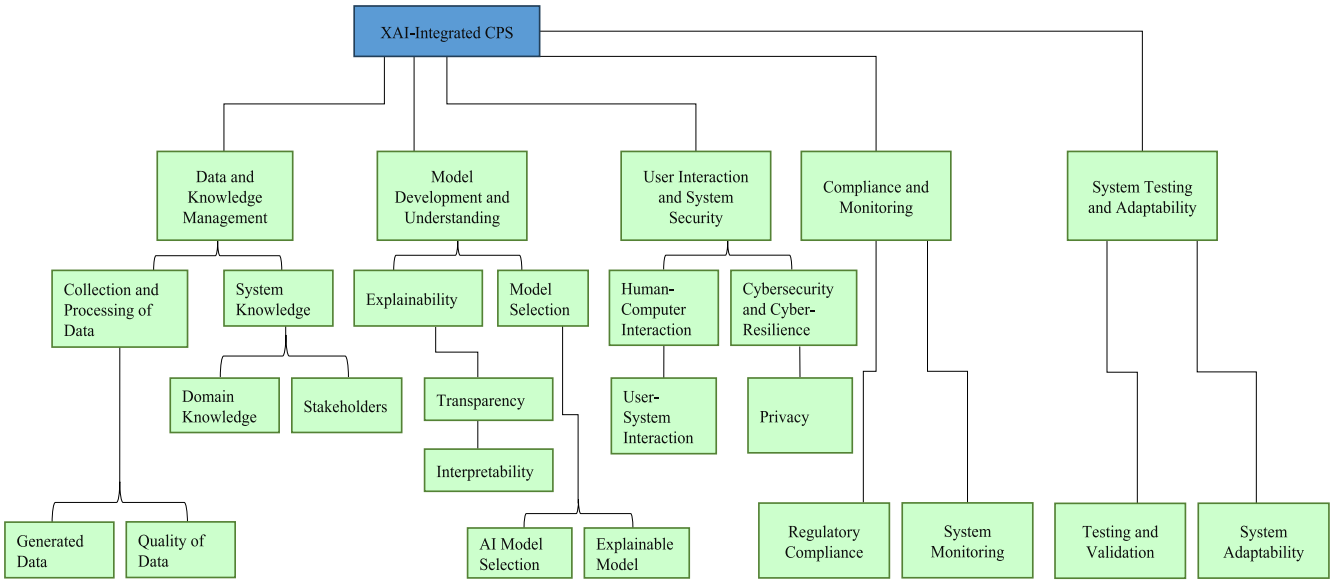

FIGURE 5. Requirements analysis diagram of an XAI-integrated cyber-physical system.

meets the business and software requirements, and provides a quality product. There are many important qualities that requirements analyses must define for successful XAI-based cyber-physical systems. Fig. 5 shows the concepts of requirements analysis for XAI-integrated cyber-physical systems.

- **Collection and Processing of Data**
  Data Used in Decision-Making: The sources and uses of data utilized in decision making must be clearly defined. Any training data must be representative, generalizable, and unbiased. Appropriate labeling methods must be selected for the training of supervised learning models.
  Generated Data: The XAI-based CPS is likely to generate large amounts of data as it functions. Plans must be made for the collection, storage, analysis, and use of these data.
  Quality of Data: The criteria for measuring data quality must be selected; evaluation metrics must be chosen. These may include accuracy, precision, recall, loss, completeness, generalizability, relevance, and others.
- **System Knowledge**
  Domain Knowledge: Define the specific domain in which the system will function; define physical processes, system requirements, and environments.
  Stakeholders: Identify and consult stakeholders to define their needs and expectations; this may include domain experts, end users, programmers, management, etc.
- **Explainability**
  Transparency: The level of transparency needed in the decision-making process should be determined. Protection of private data and trade secrets should be considered. The type of explanations and amount of detail should also be defined.
  Interpretability: Determine the level or levels of interpretability of explanations needed based on who will be using the model. Experts, users, system managers, and other stakeholders may need different levels of interpretability and detail in generated explanations.
- **Model Selection**
  AI Model Selection: The AI model or models must be chosen carefully to meet the requirements. Some considerations include the use of supervised, unsupervised, or semi-supervised learning models; generalizability; model complexity; available computational power; processing time; strengths and weaknesses of individual models; and combination of models.
  Explainable Model: If the selected AI model is not interpretable, then XAI must be added to provide explanations. Important considerations for selecting XAI models include global or local explanations, model-specific versus model-agnostic XAI, explanation type such as rule-based or feature-based, type of output such as graph-based, visual, or numeric, and others.
- **Human-Computer Interaction**
  User Interface: Develop an appropriate user interface. Consider different interfaces or outputs based on user types. Interfaces may be created based on user type or for each individual.
  User-System Interaction: Create methods for users to interact with the system, such as running queries about unexpected results, generating rules, viewing and selecting relevant features, adding or updating

rules, testing different conditions, or identifying incorrect decisions.

- **Cybersecurity and Cyber-Resilience**
  Develop and implement plans to protect the system and mitigate damages in case of cyberattack, fault states, or failures. Ensure the protection of private data including personal data, industry data, and trade secrets; protect data integrity; maintain safety in case of adversarial attack; build cyber-resilience to enable the system to continue functioning in spite of attacks or unexpected circumstances.
- **Testing and Validation**
  Define how the system will be tested to ensure that it is meeting the required standards of explainability and other measures of performance.
- **Security and Privacy of Explanations**
  Ensure that explanations will provide adequate transparency while protecting private information. Ensure an appropriate level of privacy in explanations generated for different users such as stakeholders, expert users, customers, regulatory organizations, etc.
- **Regulatory Compliance**
  Regulatory Requirements: It is necessary for any XAI-based CPS to comply with regulatory organizations, meeting the applicable legal, ethical, justification, and documentation requirements for the system's decisions. Each domain and location may have its own regulatory requirements such as laws about use of private health information, protection of trade secrets, and industry standards.
  Documentation Standards: Create and follow standards for documentation; provide adequate documentation for external audits, internal audits, performance monitoring, and system improvements.
- **System Monitoring**
  Develop methods for continuous monitoring of the system to track performance; include processes to update and improve the system as required.
- **System Adaptability**
  Ensure that the system is able to adapt to software updates, hardware updates, addition or removal of components, and maintenance. If required, define methods of testing system functionality under different circumstances such as changing environments, failure of components, changes in availability of materials, etc.

V. DISCUSSION

A. XAI: THE KEY TO FUTURE CYBER-PHYSICAL SYSTEMS

Trust and understandability are necessary for the widespread adoption of XAI and CPS. Cartolano et al. [21] state, "The problem of trust in machine learning models is critical since, without it, users will tend not to use algorithms in every field of application. Because of this, interpretable approaches become essential rather than optional." For cyber-physical systems to be widely accepted and adopted, they must not only be trustworthy but also sustainable. **Sustainability** indicates the longevity of the systems and their related infrastructure, meaning their ability to effectively fulfill their intended purpose throughout their lifespan [29]. Such systems must be able to adapt to changing environments and uncertain conditions. Weyns et al. [29] argue that cyber-physical systems must be made **smarter**, meaning that "systems and engineering processes adapt and evolve themselves through a perpetual process that continuously improves their capabilities and utility to deal with the uncertainties and amounts of data they face." These systems will continually evolve over time, learning from experience and stakeholder interaction so that they can successfully handle uncertainties, new data, and changes throughout the system lifespan. These smart systems will help CPS to be **resilient**, meaning they must be able to perform their intended functions in spite of ongoing attacks, instability, or uncertainty [29].

Explainable AI is a key part of making these research goals possible. XAI clearly explains models in a way that humans can understand, enabling them to place an appropriate level of trust in a given model and manage it effectively [29]. In addition, XAI has the function of teaching humans knowledge about explainable AI systems that they can then generalize and apply to other systems and areas.

B. THE CURRENT STATUS OF XAI FOR CYBER-PHYSICAL SYSTEMS

Explanations enhance understanding not only of the AI network, but also of the CPS itself [6]. This allows users to adapt as the CPS changes over time and to reason about possible anomalous behavior by the system. However, some researchers argue that existing XAI methods used for CPS are not effective in providing understandable explanations for time-series information [6], [70]. In addition, existing XAI methods used for CPS often are not truly understandable because they do not include contextual awareness, history of usage, and individualized user profiles [6]. Some XAI in CPS methods used to provide clearer explanations include digital staining, heatmaps, limited scope natural language, TensorFlow graph visualization, augmented reality visualizations, and immersive visualizations of explanations.

Another challenge related to CPS and XAI is the often-changing nature of CPS. Many cyber-physical systems such as smart homes, already highly customized, may have components added or removed; any XAI used in such systems must be able to take these new components into account [71]. Some new components may not have existed at the time the CPS and XAI models were developed; researchers and manufacturers must ensure that these new components can be easily integrated with the existing systems.

Although there have been many advances in the field of XAI, most existing XAI methods are designed to explain decisions made by only one agent. In cyber-physical systems, many agents may be responsible for making decisions [71].

In decentralized CPS, decision making is coordinated between many agents. Future research is required to find solutions that will meet this need.

C. ETHICS AND LEGAL CONCERNS: TRANSPARENCY, FAIRNESS, CONFIDENTIALITY, AND BIAS

Confidentiality and bias are still areas of vulnerability in XAI [13]. Many new XAI methods that are developed still use extensive training data; those data are susceptible to bias, thus leading to potential bias in a model which is supposed to be fair. In addition, the training data, testing data, or information within the model itself could be breached, leading to confidentiality risks [13]. This puts people, intellectual property, and security at risk. It is therefore crucial for developers of XAI for CPS to ensure the security of their systems. In addition, disruptions of an XAI model used in a CPS could also impede operations of that cyber-physical system. This could lead to shutdowns, equipment breaking, incorrect information being given, information being stolen, delays in business or supply chain, loss or corruption of client data, and other far-reaching consequences. Cyber-resilience goes hand-in-hand with cyber-security. Cyber-resilience can be defined as ''the ability to continuously deliver the intended outcome despite adverse cyber events'' [72]. XAI can be used to address this problem by being integrated with cyber-security to detect ongoing attacks. XAI can also be used to recommend and activate methods which allow the system to continue functioning during an attack, thus achieving cyber-resilience.

Both CPS and AI systems must be developed to include accountability. Responsible use of AI requires transparency, fairness, and accountability. Islam et al. [16] recommend that AI systems and those who implement them should be held accountable to those who are affected by the decisions they make. Because of this, stakeholders in CPS with AI must understand the necessity of XAI in high-stakes applications [13]. Cyber-physical systems require explanations for legal accountability [6].

For fair and responsible use, AI-based systems must be analyzable, convey understanding, provide transparency, and be trustworthy. Only then can their decisions be relied on for mission-critical applications [16]. XAI provides a way to meet all of these requirements. By allowing the user to understand the model, XAI also enhances trust. Understandability enables AI models to be deployed in a stable manner and trusted by experts [13].

D. MEDICAL CYBER-PHYSICAL SYSTEMS

XAI in medical CPS can be used to improve outcomes and help clinicians understand what features are most important in decision-making for a given task or classification [7]. One of the most crucial parts of designing XAI is tailoring the system using the available domain knowledge and for the target audience [16], [40].

AI provides tremendous benefits to biomedical CPS. It allows these vital systems to adapt, learn, interact with the environment, analyze the environment, find patterns, and perform complicated analyses to make predictions [7]. However, as important as AI is in biomedical CPS, the transparency and interpretability of those AI systems are equally important to performance. This is necessary to allow practitioners to understand the system behavior, recommendations, process, and errors so that they can trust the system [7]. Markus et al. [52] state that although XAI can help to ensure the trustworthiness of AI as applied to medical systems, the benefits of explainability still must be thoroughly proven through measures such as in-practice validation, extensive external validation, data quality reporting, and regulation.

E. INDUSTRIAL, SMART CITY, AND ENVIRONMENTAL APPLICATIONS

According to Sofianidis et al. [17], current implementations of AI in manufacturing do not use the full capabilities of ML and DL. Instead, they are focused primarily on data consolidation and data analysis for use with relatively simple concepts such as predictive maintenance or industrial simulations. However, ''real-life manufacturing environments are complex, dynamic and unpredictable, which highlights safety, reliability and trustworthiness challenges for the respective AI deployments'' [17]. Advanced AI for CPS in manufacturing must address the issues of transparency, explainability, development of successful ways for the AI to interact with the manufacturing environment, implementation of human-centric AI, protection against cybersecurity risks, and possibly inaccurate or unreliable industrial data. Reliability, security, and safety are paramount in complex AI-integrated CPS in manufacturing. Sofianidis et al. [17] remark, ''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.''

Sofianidis et al. [17] state that AI can create remarkable changes in manufacturing by improving event prediction, fault diagnosis, safety, quality inspection, production planning, production management, efficiency, process monitoring, and many others. AI applied to manufacturing is a major driver of better quality. However, lack of trust in the AI model due to opacity hinders the decision-making process [17]. It is necessary to augment or exchange AI models with XAI to enable understanding, trust, and optimal decision making. XAI can be applied to industrial CPS in many ways. For example, XAI can be used with existing data from CPS devices to analyze and predict operational management of smart industrial machines, thus optimizing time and resources [53].

XAI is a fundamental component of smart cities; any AI-based decisions must be trustworthy and yield justifiable explanations. Houda et al. [62] recommend that XAI-based

intrusion detection systems for smart city applications should provide specific explanations for both model users and security experts. The authors demonstrate the importance of providing both local and global explanations in this context. Global explanations can be used to determine the most important features which demarcate the presence or absence of an intrusion, for example. Local explanations are necessary to provide information about specific instances in question, such as in the case of a suspected attack or an instance marked as an intrusion but which may in fact be normal traffic [62].

Vital smart city infrastructure such as smart power grids, smart transportation, or smart water management systems must be not only trustworthy and explainable but also robust; XAI increases the robustness of these services [34]. In agriculture and food production CPS, sensors can be integrated with irrigation systems to monitor temperature and humidity. A case discussed by Jagatheesaperumal et al. [53] uses an explainable decision support system with fuzzy rule-based automation to increase interpretability and trustworthiness.

In CPS integration in the chemical industry, for example, it is necessary for the CPS to include methods for system control and optimization, two of the most crucial topics in the industry [5]. Conventional process control tools are not sufficient to meet the needs of advanced CPS. Cyber-physical systems are complicated and dynamic, and Industry 4.0 demands both accurate forecasting and rapid flexibility.

Combining XAI with advanced information and communication technologies supports the development of smart technology in cyber-physical systems [13]. Oliveira et al. [5] provide the example of AI and CPS in the chemical industry. This domain could save billions of dollars by implementing cyber-physical systems that help to prevent costly human errors [5]. However, for the industry to reach this level of savings, it is necessary for research to validate such large-scale CPS and AI operations. Crucially, this would require an interdisciplinary approach to demonstrate the value of such technologies [5]. Difficulty in sharing domain-specific knowledge is a great hindrance to achieving global, multi-industry understanding of this concept. XAI can be implemented to demonstrate the savings CPS can offer in a way that is understandable to stakeholders from different fields, levels of expertise, cultures, and languages.

XAI can be added to legacy equipment to improve its functioning, offer solutions for troubleshooting, and provide explainability. Eramo et al. [59] discuss the application of a smart platform entitled Prodevelop for real-time monitoring of port activities. The system analyzes data from IoT sensors, legacy information systems, and external information systems, rapidly assessing large amounts of data through cloud systems or virtual machines. We posit that XAI may be added to such systems to facilitate the planned mitigation of bottlenecks, errors, and other problems with information flow by identifying or altering rules that determine system operation and information flow, identifying characteristics of all nodes and pathways which exhibit bottlenecks, improving anomaly detection, and increasing cyber-resilience. Jagatheesaperumal et al. [53] remark, ''XAI models integrated with meta-learning strategies are largely used in cyber-physical systems that are the core components of Industry 4.0. They ensure rich simulation infrastructure, smart communication with machines, higher level of visualization, better analysis of service quality and maximization of production efficiency.''

Given the in-real-time nature of the project in Eramo et al. [59], explainable AI can facilitate live monitoring of the system by quickly providing human-understandable explanations. Thus, users or system managers can detect, understand, and take action to resolve problems as they occur. XAI can also generate explanations for queries and propose possible solutions with explanations to justify them. XAI can be added to legacy systems so that old equipment does not need to be replaced, saving time and money. Machine learning or deep learning can be added to legacy CPS to provide analysis, support, recommendations, and greater efficiency. In systems with existing and effective ML or DL, analysis may be performed to determine which XAI models suit the particular ML or DL algorithm in use. System developers can evaluate what kinds of explanations are needed, select whether to use a model-specific or model-agnostic type of XAI, and choose the specific model that is needed.

XAI can also be deployed to assess cost-effectiveness, model accuracy, and model quality [53]. In cities with legacy equipment that are transitioning to smart cities, XAI can offer adaptable solutions to ''decisions made by the smart IoT-based devices employed for the growth of smart cities'' [53], yielding cities that are smarter, sustainable, and efficient. To make these adaptations possible, further research is required to ensure smooth development and integration in areas such as security, resource management, and control. In addition, explainable AI may be employed in existing cyber-physical systems to advance automatic learning for predictive maintenance in cyber-physical production systems, product quality improvement, development of robots or cobots to perform accurate, effective remote quality checks, and crucially, adaptation of systems to meet growing levels of demand. XAI can be applied to existing health monitoring systems such as pandemic monitoring to improve analysis and prediction [13].

F. VULNERABILITIES OF CPS-BASED CRITICAL INFRASTRUCTURE

As an integral component of modern critical infrastructure, CPS are especially vulnerable to attack [1], [14]. AI is used to monitor and control life-sustaining cyber-physical systems such as smart power grids or water management systems. Such CPS infrastructures are vulnerable to cyberattacks that can include viruses, access by unauthorized users, or altering, interception, or deletion of information [14]. Because of this, machine learning is increasingly used in critical CPS applications. Understanding the rationale behind decisions

made by these systems is necessary, now even required by some governments [14].

G. CASE STUDY EXAMPLE: SMART WATER TREATMENT SYSTEM

To illustrate implementation of XAI in a real-world cyber-physical system, we present a hypothetical case study of a smart water treatment system. This CPS is responsible for many tasks such as monitoring, intaking, sensing, transferring, administering chemicals to, filtering, and testing water. AI can be used to automate system actions, provide control, ensure water quality, and provide cybersecurity. XAI can be applied to this to enhance security, provide cyber-resilience, provide more knowledge about the system, and prepare for contingencies. The smart water filtration system is susceptible to physical attacks which affect the behavior of the physical parts of the system and attacks on cyber components which alter the behavior of the computing elements. For example, an adversarial attack could cause the system to indicate that unsafe water is safe to drink by altering data within the system or by modifying the behavior of the chemical sensors. Cyberattacks could also affect the behavior of actuators and cause dangerous levels of treatment chemicals to be added into the water.

XAI can be used to monitor for such attacks, identify features that are predictive of the presence and type of attack, and provide recommendations for actions to mitigate the problem. Explainable AI can also be used to query the system to see how it is functioning, what rules it is operating under, and how each area is performing. Virtual reality can be added to show system components, model water and chemical flow, and allow for queries at key locations such as sensors, pipes, water tanks, or actuators. The explainable AI model can provide information about these components such as component status, feature values, or water flow. Such an interface would provide a more intuitive way to view the system, providing deeper knowledge and allowing users to see into the system in ways that were not possible before. A visual display using colors can alert users to problems or anomalous values. In addition, the XAI model can find relevant features of attacks, discover new predictive features, and develop knowledge or rules to improve the functioning of the system. These rules and concepts can be verified by experts, providing for a safer and more efficient water treatment system while also creating transferable knowledge that can be used to improve other systems.

Another way in which XAI may be helpful is in prediction, early detection, and mitigation of unsafe water quality such as that caused by harmful algal blooms. The XAI interface can display features of the water and system over time as well as environmental features such as weather, conditions in the water source, time and date, pollution, precipitation, health of organisms in the water source, types and presence of micro-organisms, and any other available features which may be relevant. XAI can be applied to identify features predictive of harmful algal blooms over time. The XAI can then provide human-understandable explanations of what the relevant features are, when the features reached levels of interest, and how they changed over time. This can create a framework for prediction, early detection, and protective measures for changes in water quality due to harmful algal blooms and other occurrences.

H. CASE STUDY EXAMPLE: AI-INTEGRATED CCTV SURVEILLANCE

Some smart cities are now implementing AI-integrated CCTV (closed-circuit television) for monitoring of criminals or prevention of crime. In this case, video footage is analyzed by computer vision and artificial intelligence to produce classifications or decisions. This application is sensitive, placing the reputations and lives of people at stake, especially if the decisions made by the AI system are wrong. Javed et al. [34] state that ''Traditional computer vision applications do not explain or justify the classification of images/videos. Hence, making real-time decisions based on the classifications given by computer vision-based applications in scenarios in smart cities, such as collision avoidance, traffic monitoring and crime prevention, may incur severe costs, such as loss of lives and ethical issues.'' In addition, there is a significant risk that such technologies will be subject to existing biases in criminal justice systems, especially if the AI models are trained on biased datasets. Taylor and Taylor [27] remark on AI, ''commercial models that are trained on biased data sets will treat underrepresented groups unfairly and inaccurately.'' It is therefore imperative that if CCTV is to be used in crime detection and prevention, special care must be taken to ensure the system is fair and unbiased. XAI can aid in achieving this goal by providing explainability, justification of classifications, and interpretability [34]. In addition, preventative measures may be a beneficial use for these systems. Explainable AI can identify relevant features such as behavioral indicators which may show intent to commit a crime; once this is identified, preventative measures can be implemented to deter the crime from occurring, such as deploying extra security officials to a place of interest at a shopping mall. Locations, times, dates, valuable assets, and other features can be analyzed using AI, then explored and explained via XAI to determine where and when crime is likely to occur. Then, preventative measures such as increased security or safe storage of valuable objects can be implemented to decrease the likelihood of the potential crime.

I. SECURITY IN CPS

Cyberattacks on cyber-physical systems can lead to theft or corruption of sensitive data, system failure, user harm, and alteration or harm to the physical system and its operations [41]. Security and privacy are therefore essential in CPS design and deployment. Challenges in CPS security include unexpected events, rapid changes to the environment, and the risk of high false alarm rates degrading threat recognition and system functioning. As the CPS becomes larger and its interconnections become increasingly complicated, more

opportunities for attack are created. It also becomes more difficult to correctly identify threats in very large systems [41]. XAI can improve cybersecurity by explaining the reasons that attacks were misclassified as normal traffic [14], and vice versa. This allows users to understand what happened and what measures should be taken to solve the issues, leading to effective diagnosis, debugging, and attack prevention. Marino et al. [14] say of intrusion detection systems, "It is crucial that the inner workings of data-driven models are transparent for the engineers designing IDSs."

In intrusion detection systems for cyber-physical systems and IoT networks, ML/DL-based systems make machine-centric decisions which may be executed by human users such as high-level cybersecurity staff [73]. Usually, these systems are opaque, meaning they are not inherently understandable by human users and they do not provide explanations as to how their decisions are made. Without transparency, the results from the MD/DL model cannot be properly understood and therefore cannot be used optimally by humans to protect and maintain optimal functioning of the system. The authors highlight the importance of using XAI as a solution to this conundrum, offering models such as RuleFit and SHAP to add trust and transparency as well as explanation [73].

Another important future direction for research is the development of more unsupervised XAI methods. Much of the data generated by cyber-physical systems is unlabeled [1]. To take advantage of newly generated data and improve ever-changing systems without time-consuming, expensive manual labeling, unsupervised XAI will be necessary. Unsupervised XAI can detect possible anomalous behaviors without knowing what they are beforehand using methods such as self-organizing maps and clustering. XAI can provide a great advantage in cybersecurity given the ever-developing nature of cyberattacks [1], [14].

Cybersecurity and cyber-resilience for smart transportation are crucial for maintaining safety. Cyberattacks which lead to accidents could cause loss of materials, loss of life, and environmental damage or hazardous conditions when targeted at hazardous materials transportation [42]. To mitigate these dangers, smart XAI solutions must be implemented so that stakeholders can both trust the system and be a part of the process of ensuring cybersecurity and cyber-resilience.

VI. RECOMMENDATIONS

A. EVALUATION METRICS AND FORMALIZED STANDARDS

One of the key conclusions of this study is that the field of XAI for cyber-physical systems has relatively few quantitative studies. The majority of the literature focuses on what a new model is meant to do, not quantitative analysis of how well it has been done. Thorough evaluation metrics are rarely provided in these studies; however, these would make it easier to compare studies and determine how well a new method performed. Quantitative metrics would provide numeric measures of how well a system performs, show areas for improvement, allow numeric comparison between different methods, and highlight specific characteristics of the model's performance such as accuracy, F1 score, loss, recall, and precision. Another concern is that many of the academic research articles which do provide demonstrations of methods for XAI in cyber-physical systems have run simulations but have not tested or implemented their methods in real-world scenarios.

According to Islam et al. [16], it is necessary in future research to develop a formalized, generic framework for XAI. This framework should include providing formalized explanations, giving customized explanations for different users or recipients (for example, people of different fields or levels of expertise, laypersons, or other machines), developing methods of quantifying the generated explanations, and finding ways to quantify the human comprehensibility of explanations. Combining knowledge from multiple domains can enable the development of such a framework. As XAI models that comply with such a framework become available, even mission-critical or other high-stakes CPS from different fields would be enabled to apply XAI methods to their black box models to achieve greater confidence in their results and comply with legal, ethical, and organizational regulations [16]. A formalized framework would also allow legal, ethical, and regulatory bodies know what to expect from the systems and have a clearly defined, standardized way to measure their performance.

Standardized vocabulary and definitions are needed in the field of XAI to facilitate the sharing and comparison of information [26]. The criteria that enable trust in systems must be clearly defined and formalized. Presently existing criteria are ambiguous, needing to be broken down into smaller, definable, and testable constructs. Similarly, there are not enough empirical studies to clearly define and validate user-based measures of interpretability [26]. Most of the presently developed XAI methods depend on the intuition of the developers for what may be explainable to humans, rather than developing measures which are tested and proven to be explainable [63]. Clear definitions are required to precisely define what explainability is and how it can be measured.

B. EXPERT SUPERVISION AND LEARNING FROM XAI EXPLANATIONS

Predictions made by XAI are still subject to flaws, bias, errors, and inconsistencies [13]. Sometimes those predictions may be unreasonable or out of the correct sequence. Users need to understand that the decisions made by the XAI model may still be incorrect, even if it sounds reasonable given the explanation [27]. Because of the risk of incorrect decisions, it is important that XAI models allow for experts to compare their knowledge with the knowledge learned by the algorithm [21]. This can help the experts to detect incorrect rules or recommendations in the model. It can also help experts to learn more from the XAI itself. DL in particular is a powerful way to find even slight relationships between

variables [27]. Users can examine the model's learned rules or features and assess whether or not they are correct using their expert knowledge.

C. INTERDISCIPLINARY COLLABORATION AND COMMUNICATION

XAI provides a solution to the problem of interdisciplinary communication in the grand challenges facing future cyber-physical system development. According to Rajkumar et al. [11], "Many grand challenges await in the economically vital domains of transportation, healthcare, manufacturing, agriculture, energy, defense, aerospace and buildings. The design, construction and verification of cyber-physical systems pose a multitude of technical challenges that must be addressed by a cross-disciplinary community of researchers and educators." Well-designed XAI makes explanations and understanding readily understandable to people from different fields, cultures, and levels of expertise.

Cyber-physical systems combine informatics, computer science, automation, and engineering [5], [74]. A transdisciplinary approach is needed for CPS research but does not yet exist [5]. Rajkumar et al. [11] remark that the discipline of CPS will also require integration of computing system theories, communication systems theories, human-CPS interaction, and physical system control and sensing. Physics, chemistry, language, system architecture, hardware, programming, human factors, mathematics, robustness, and cybersecurity must all be taken into account to develop optimal CPS [11]. Appropriate, comprehensible explanations from XAI-integrated CPS will enable people from all of these domains to understand, learn from, interact with, and change the systems as a collaborative, interdisciplinary effort.

Experts from many fields will be needed to create the XAI-integrated cyber-physical systems of the future. NLP experts will be needed to ensure proper generation of natural language explanations [59]. Data scientists and statisticians can collaborate to develop formalized statistical measures and intuitive scientific visualizations. Together, machine learning experts and social scientists can define the specific requirements needed for explainability in anomaly detection systems for CPS [63]. Psychologists and human factors experts can test and verify trust-inducing criteria. Psychologists can also determine which kinds of explanations are most effective in different scenarios [31]. Richer loss functions must be developed for real-world situations with a greater number of performance criteria [26]. Such loss functions are especially needed for cyber-physical systems, which are often complex with many elements, each with its own goals and criteria.

Human factors and psychology can work synergistically to develop formalized ways to measure the human comprehensibility of explanations. These fields are also necessary to measure the effectiveness of explanations when shown to people of different levels of expertise and different domains. This is of critical importance in any application. For example, in medical CPS which provide diagnostics, outcome predictions, and treatment recommendations, practitioners need understandable explanations not only for best practice but also for ethical, legal, and documentation reasons. In smart agriculture, farmers and agronomists will need explanations they can understand, trust, and implement. In both examples, the users will need to be able to compare their expertise with the recommendations from the system to learn from it, assess whether or not the current model is trustworthy, and make sure the results are correct.

D. CUSTOMIZABLE OUTPUT FOR DIFFERENT USERS

Understandable visual output is key in fields or applications where the final users will not be ML experts, such as auditors, maintainers, or managers [6], [21]. Without understanding and trust, users from any field will tend not to use ML [21]. Developing user profiles and providing tailored explanations to each type of user will facilitate understandability [6]. To customize explanations for each user, Darias et al. [75] developed a method using Case-Based Reasoning to select the best explanation method for each situation based on the AI model, domain, and user preferences. Customized explanations enable users to learn more from the explanations and gain a better understanding of the CPS.

E. MULTIMODAL EXPLANATIONS

A key part of future development in XAI in CPS may be multimodal or even multisensory explanations that provide information in different ways. Zizic et al. [38] state that operators in future industry "should collaborate with the equipment by using [their] own physical, sensorial, and cognitive capabilities in an environment that provides safe work and technological assistance" and that real-time information should be provided. We suggest that multisensory output may be useful to operators, particularly those with disabilities, as another explanation method to quickly understand what is going on in the system. For example, researchers are developing augmented or extended reality explanation methods, such as displaying an industrial shop floor, to provide explanations for CPS [6]. We recommend that researchers work creatively to find new, unexplored methods of providing explanations, new XAI models, and ways to customize output to suit the needs of different users.

VII. FUTURE RESEARCH DIRECTIONS AND RECOMMENDATIONS

A. INDUSTRY 5.0

Explainable artificial intelligence will be a critical element in the proposed Fifth Industrial Revolution (Industry 5.0). The goal of the Fifth Industrial Revolution is to develop future industry to build prosperity that is integrated into both work and everyday living [76]. This will be achieved in part through intelligent and cutting-edge technologies like XAI. While the Fourth Industrial Revolution has improved human-machine

interactions in terms of technical functioning, Industry 4.0 has not focused on social sustainability. Industry 5.0 extends the ideas of Industry 4.0 to include environmental and social sustainability [38], allowing humans to use their skills, abilities, knowledge, and expertise in collaboration with machines and robots. Industry 5.0 focuses on a human-centric approach, sustainability, and resilience, emphasizing reduction of environmental impact and creation of flexible production processes.

Integrating XAI with current technology and CPS will facilitate collaborative relationships between machine/robot and human. Rather than replacing the human entirely or keeping the human unaware of what is going on within the AI model as in a traditional black box model, XAI enables the human to understand, learn from, manage, and interact with the model. This can designate the human as the expert with the final say, focusing on their knowledge that can monitor, problem solve, remove bias from, create better rules for, or evaluate a system. XAI will be a crucial part of building Industry 5.0 in a human-centric way. As Zizic et al. [38] state, ''The ideal type of the factory worker of the future is participative and proactive.'' XAI enables the workers to participate, understand, and be proactive in CPS processes by granting them understanding and agency to instate changes. Zizic et al. [38] also emphasize that human workers need to be able to develop their skills and creativity. XAI can make them a partner in the process, allowing humans to implement their creativity by coming up with new ideas and solutions for the model.

XAI enables CPS to have better resilience. As one of the three key Industry 5.0 goals, resilience entails flexibility, adaptable production capacities, and the ability to continue functioning in a crisis. Rather than creating an AI system, optimizing it, and implementing it to run on its own in a set manner, XAI makes the system more flexible and responsive. XAI allows for testing and makes the system interactive; it can be designed with the ability to change readily. This makes it easier for humans to test the system with many different rules and scenarios. The greater control and understanding of a CPS that XAI can give makes it possible for humans to develop in advance a flexible, resilient system that can be altered to meet different needs, even in a crisis situation. This is especially important for CPS such as production systems which require large amounts of physical resources and will be broadly affected by any disruptions. Testing diverse situations in advance using XAI will allow users to plan ahead and mitigate the effects of unexpected crises.

B. WORK ENGAGEMENT AND PREVENTION OF OCCUPATIONAL BURNOUT

The human-centric research topics for Industry 5.0 include human-centered integration, work engagement, and occupational burnout [38]. We recommend that XAI may be used as one solution to assist in preventing burnout and facilitating work engagement. According to Bakker et al. [77], ''Whereas burnout refers to a state of exhaustion and cynicism toward work, engagement is defined as a positive motivational state of vigor, dedication, and absorption.'' When employees feel cynical and fatigued about their work, they are more likely to experience significant physical and mental health problems. Over time, a higher level of burnout has been correlated with increased likelihood of infections, type 2 diabetes, musculoskeletal disorders, and cardiovascular diseases [77]. Occupational burnout and employee engagement have a profound effect on individuals and organizations as a whole.

XAI can allow employees to exercise some control over the AI-based CPS, engage in the decision making process, feel a sense of meaning in their work, and take part in opportunities for personal growth and learning. This may grant workers a greater sense of control, autonomy, task significance, and task variety; all of these are predictors of work engagement [77]. Providing employees with opportunities for personal development and a variety of tasks may reduce cynicism, one of the two main elements of occupational burnout. It is in the best interest of organizations to address these concerns as burnout and work engagement explain a significant portion of the variation in organizational behavior [77].

XAI facilitates human engagement with the AI-based cyber-physical system. Instead of having a fully autonomous system that makes decisions for humans and requires them to do repetitive tasks, XAI enables the human to interact with the system, make decisions, use their knowledge and expertise, learn about the topic, and have a final choice where appropriate. This gives the human a sense of agency, control, and the option for creativity. In addition, a once-meaningless task now becomes meaningful as the user can better understand the information given and the inner workings of the system. We suggest that implementation of XAI in CPS will help to decrease employee burnout and increase work engagement, creating a positive impact at the individual level and for the entire organization.

C. MULTIPLE CPS INTEGRATION AND SYSTEM COGNITION

One of the key areas of future development in XAI and CPS is the combination of multiple cyber-physical systems in a way that enables integration, intercommunication, and cooperation for autonomous task completion [5]. These may be developed into a cognitive system that requires little human supervision or intervention. Cognition ''allows the modelling, representation and learning of complex behaviors and interactions between the system components and the system data'' [5]. Cognition can be achieved with supervised or unsupervised learning. It enables the AI models to learn from the system over time, creating adaptability for the cyber-physical system.

D. SMARTER CPS: SELF-EXPLAINABILITY AND CONTEXT AWARENESS

One of the most important future directions of cyber-physical systems is to make them smarter by creating self-explainability, smarter ecosystems, and assurances for

unknowns [29]. As smart cyber-physical systems will use artificial intelligence to make decisions, the authors recommend including XAI in those systems to provide explanations of the rationale behind their decisions when queried. Jha [6] states, ''To become self-explainable, a system would need to know its working environment, internal states, user profiles, and the interactions between its software and physical components.'' The author recommends providing context-aware explanations as a critical enhancement to enable the effective application of XAI to CPS. Context-aware explanation systems would allow users to have a better understanding of the behavior of the CPS in unanticipated scenarios such as in a new location or under different working conditions [6]. The constantly evolving context of CPS must be included and modeled in context-aware explanations.

VIII. CONCLUSION

In this research, we have reviewed the application of XAI to cyber-physical systems. We have defined each area and demonstrated the necessity of integrating XAI into CPS, showcasing the benefits, challenges, importance, and future research directions. We provided a detailed requirements analysis for development of XAI-based cyber-physical systems. We have contributed new recommendations for future systems and detailed discussions of the multidisciplinary goals and benefits of XAI for CPS.

Applying XAI to CPS ensures explainability for AI-based choices that affect a multitude of outcomes such as manufacturing processes, resource usage, safety, security, medical care, people, and the environment. It can also be used to boost protection against many of the vulnerabilities facing cyber-physical systems. These improvements in CPS can lead to greater efficiency, less waste, more stable industries and economies, and greater social and environmental sustainability.

XAI can also encourage implementation of AI in CPS, enabling stakeholders to gain the greatest benefits from artificial intelligence. Without an understanding of how AI works, many industries and other entities are hesitant to adopt artificial intelligence [7], [29]. If it can be demonstrated to stakeholders that XAI increases system trustworthiness and control, they may be more willing to implement AI in their cyber-physical systems.

Key gaps in the current literature include a dearth of real-world experimental implementations of newly developed XAI models for CPS, a lack of detailed evaluation metrics in such studies, the necessity of standardized vocabulary and definitions, and the need for comparative evaluation techniques. Formalized techniques to measure the human understandability of explanations are also urgently needed [16]. There is also a need for benchmark tests designed specifically to analyze and compare the performance of XAI models for CPS.

Interdisciplinary collaboration is necessary to create optimized XAI-enabled cyber-physical systems, combining knowledge from fields such as data science, natural language processing, human factors, psychology, and statistics to ensure effective and understandable XAI. Explanations must be suited for different types of users depending on their level of expertise, familiarity, and field of study. When users are able to understand the reasoning behind the decisions or recommendations made by the model, they can learn from it, add new rules, monitor for errors, and make modifications to build a stronger system.

XAI paired with cyber-physical systems has the ability to revolutionize current and future industry, infrastructure, and other smart technologies. It provides users with the opportunity to understand, learn from, monitor, analyze, correct, and test the system. This creates better working conditions, meets Industry 5.0 goals, and puts humans in charge of the artificial intelligence. We recommend that application of XAI to CPS may decrease burnout by increasing employee engagement, giving the human in the loop agency, synergistically developing expertise, and providing opportunities for creativity. We also advocate the development of multisensory outputs and explanations in XAI systems.

From smart infrastructure to smart transportation and beyond, XAI-based CPS enables better use of existing resources and responsible use of AI. It improves user understanding of cyber-physical systems, works synergistically with expert knowledge to provide the best decisions, can readily be altered or tested with different rules, and can be utilized to teach users more about the given topic, thus adding to the body of scientific knowledge. XAI for CPS makes possible interdisciplinary collaboration by providing clear explanations to users of different levels of expertise, creating transferable knowledge that CPS experts can harness to create the technology of the future.

ACKNOWLEDGMENT

The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the U.S. Government.

REFERENCES

[1] C. S. Wickramasinghe, K. Amarasinghe, D. L. Marino, C. Rieger, and M. Manic, ''Explainable unsupervised machine learning for cyber-physical systems,'' *IEEE Access*, vol. 9, pp. 131824–131843, 2021. [Online]. Available: https://ieeexplore.ieee.org/document/9536751

[2] R. Alguliyev, Y. Imamverdiyev, and L. Sukhostat, ''Cyber-physical systems and their security issues,'' *Comput. Ind.*, vol. 100, pp. 212–223, Sep. 2018, doi: 10.1016/j.compind.2018.04.017.

[3] F. Hu, Y. Lu, A. V. Vasilakos, Q. Hao, R. Ma, Y. Patil, T. Zhang, J. Lu, X. Li, and N. N. Xiong, ''Robust cyber-physical systems: Concept, models, and implementation,'' *Future Gener. Comput. Syst.*, vol. 56, pp. 449–475, Mar. 2016, doi: 10.1016/j.future.2015.06.006.

[4] J. Lee, B. Bagheri, and H.-A. Kao, ''A cyber-physical systems architecture for Industry 4.0-based manufacturing systems,'' *Manuf. Lett.*, vol. 3, pp. 18–23, Jan. 2015, doi: 10.1016/j.mfglet.2014.12.001.

[5] L. M. C. Oliveira, R. Dias, C. M. Rebello, M. A. F. Martins, A. E. Rodrigues, A. M. Ribeiro, and I. B. R. Nogueira, ''Artificial intelligence and cyber-physical systems: A review and perspectives for the future in the chemical industry,'' *AI*, vol. 2, no. 3, pp. 429–443, Sep. 2021. [Online]. Available: https://www.mdpi.com/2673-2688/2/3/27

[6] S. S. Jha, "An overview on the explainability of cyber-physical systems," in *Proc. Int. FLAIRS Conf.*, vol. 35, 2022, pp. 1–4.
[7] N. Alimonda, L. Guidotto, L. Malandri, F. Mercorio, M. Mezzanzanica, and G. Tosi, "A survey on XAI for cyber physical systems in medicine," in *Proc. IEEE Int. Conf. Metrol. Extended Reality, Artif. Intell. Neural Eng. (MetroXRAINE)*, Oct. 2022, pp. 265–270. [Online]. Available: https://ieeexplore.ieee.org/document/9967673
[8] S. Sivamohan, S. S. Sridhar, and S. Krishnaveni, "TEA-EKHO-IDS: An intrusion detection system for industrial CPS with trustworthy explainable AI and enhanced krill herd optimization," *Peer-to-Peer Netw. Appl.*, vol. 16, no. 4, pp. 1993–2021, Aug. 2023, doi: 10.1007/s12083-023-01507-8.
[9] L. Almuqren, M. S. Maashi, M. Alamgeer, H. Mohsen, M. A. Hamza, and A. A. Abdelmageed, "Explainable artificial intelligence enabled intrusion detection technique for secure cyber-physical systems," *Appl. Sci.*, vol. 13, no. 5, p. 3081, Feb. 2023. [Online]. Available: https://www.mdpi.com/2076-3417/13/5/3081
[10] A. P. Patil, J. Devarakonda, M. Singuru, S. Tilak, and S. Jadon, "XAI for securing cyber physical systems," in *Proc. 3rd Int. Conf. Secure Cyber Comput. Commun. (ICSCCC)*, May 2023, pp. 671–677.
[11] R. Rajkumar, I. Lee, L. Sha, and J. Stankovic, "Cyber-physical systems: The next computing revolution," in *Proc. Design Autom. Conf.*, Jun. 2010, pp. 731–736, doi: 10.1145/1837274.1837461.
[12] A. Rai, "Explainable AI: From black box to glass box," *J. Acad. Marketing Sci.*, vol. 48, no. 1, pp. 137–141, Jan. 2020, doi: 10.1007/s11747-019-00710-5.
[13] I. Ahmed, G. Jeon, and F. Piccialli, "From artificial intelligence to explainable artificial intelligence in Industry 4.0: A survey on what, how, and where," *IEEE Trans. Ind. Informat.*, vol. 18, no. 8, pp. 5031–5042, Aug. 2022.
[14] D. L. Marino, C. S. Wickramasinghe, and M. Manic, "An adversarial approach for explainable AI in intrusion detection systems," in *Proc. 44th Annu. Conf. IEEE Ind. Electron. Soc.*, Oct. 2018, pp. 3237–3243.
[15] A. M. Roth, J. Liang, and D. Manocha, "XAI-N: Sensor-based robot navigation using expert policies and decision trees," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Sep. 2021, pp. 2053–2060. [Online]. Available: https://ieeexplore.ieee.org/document/9636759/
[16] S. R. Islam, W. Eberle, S. K. Ghafoor, and M. Ahmed, "Explainable artificial intelligence approaches: A survey," 2021, *arXiv:2101.09429*.
[17] G. Sofianidis, J. M. Rožanec, D. Mladenić, and D. Kyriazis, "A review of explainable artificial intelligence in manufacturing," in *Trusted Artificial Intelligence in Manufacturing: A Review of the Emerging Wave of Ethical and Human Centric AI Technologies for Smart Production*, J. Soldatos and D. Kyriazis, Eds. Boston, MA, USA: Now, 2021, ch. 5, pp. 93–113.
[18] B. Pradhan, S. Lee, A. Dikshit, and H. Kim, "Spatial flood susceptibility mapping using an explainable artificial intelligence (XAI) model," *Geosci. Frontiers*, vol. 14, no. 6, Nov. 2023, Art. no. 101625, doi: 10.1016/j.gsf.2023.101625.
[19] A. Dikshit and B. Pradhan, "Interpretable and explainable AI (XAI) model for spatial drought prediction," *Sci. Total Environ.*, vol. 801, Dec. 2021, Art. no. 149797.
[20] A. Abdollahi and B. Pradhan, "Explainable artificial intelligence (XAI) for interpreting the contributing factors feed into the wildfire susceptibility prediction model," *Sci. Total Environ.*, vol. 879, Jun. 2023, Art. no. 163004, doi: 10.1016/j.scitotenv.2023.163004.
[21] A. Cartolano, A. Cuzzocrea, G. Pilato, and G. M. Grasso, "Explainable AI at work! What can it do for smart agriculture?" in *Proc. IEEE 8th Int. Conf. Multimedia Big Data (BigMM)*, Dec. 2022, pp. 87–93. [Online]. Available: https://ieeexplore.ieee.org/document/9999114
[22] N. Almakayeel, S. Desai, S. Alghamdi, and M. R. N. M. Qureshi, "Smart agent system for cyber nano-manufacturing in Industry 4.0," *Appl. Sci.*, vol. 12, no. 12, p. 6143, Jun. 2022, doi: 10.3390/app12126143.
[23] H. Elhoone, T. Zhang, M. Anwar, and S. Desai, "Cyber-based design for additive manufacturing using artificial neural networks for Industry 4.0," *Int. J. Prod. Res.*, vol. 58, no. 9, pp. 2841–2861, May 2020, doi: 10.1080/00207543.2019.1671627.
[24] M. Ogunsanya and S. Desai, "Physics-based and data-driven modeling for biomanufacturing 4.0," *Manuf. Lett.*, vol. 36, pp. 91–95, Jul. 2023.
[25] K. Alikhademi, E. Drobina, D. Prioleau, B. Richardson, D. Purves, and J. E. Gilbert, "A review of predictive policing from the perspective of fairness," *Artif. Intell. Law*, vol. 30, no. 1, pp. 1–17, Mar. 2022, doi: 10.1007/s10506-021-09286-4.
[26] F. K. Došilović, M. Brcic, and N. Hlupic, "Explainable artificial intelligence: A survey," in *Proc. 41st Int. Conv. Inf. Commun. Technol., Electron. Microelectron. (MIPRO)*, May 2018, pp. 0210–0215. [Online]. Available: https://ieeexplore.ieee.org/document/8400040
[27] J. E. T. Taylor and G. W. Taylor, "Artificial cognition: How experimental psychology can help generate explainable artificial intelligence," *Psychonomic Bull. Rev.*, vol. 28, no. 2, pp. 454–475, Apr. 2021, doi: 10.3758/s13423-020-01825-5.
[28] A. Ferraro, A. Galli, V. Moscato, and G. Sperlì, "Evaluating eXplainable artificial intelligence tools for hard disk drive predictive maintenance," *Artif. Intell. Rev.*, vol. 56, no. 7, pp. 7279–7314, Jul. 2023, doi: 10.1007/s10462-022-10354-7.
[29] D. Weyns, J. Andersson, M. Caporuscio, F. Flammini, A. Kerren, and W. Löwe, "A research agenda for smarter cyber-physical systems," *J. Integr. Design Process Sci.*, vol. 25, no. 2, pp. 27–47, Aug. 2021. [Online]. Available: https://content.iospress.com/articles/journal-of-integrated-design-and-process-science/jid210010
[30] A. B. Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. Garcia, S. Gil-Lopez, D. Molina, R. Benjamins, R. Chatila, and F. Herrera, "Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI," *Inf. Fusion*, vol. 58, pp. 82–115, Jun. 2020. [Online]. Available: https://www.sciencedirect.com/science/article/abs/pii/S1566253519308103
[31] T. Speith, "A review of taxonomies of explainable artificial intelligence (XAI) methods," in *Proc. ACM Conf. Fairness, Accountability, Transparency*, Jun. 2022, pp. 2239–2250, doi: 10.1145/3531146.3534639.
[32] N. Capuano, G. Fenza, V. Loia, and C. Stanzione, "Explainable artificial intelligence in CyberSecurity: A survey," *IEEE Access*, vol. 10, pp. 93575–93600, 2022.
[33] G. Yang, Q. Ye, and J. Xia, "Unbox the black-box for the medical explainable AI via multi-modal and multi-centre data fusion: A mini-review, two showcases and beyond," *Inf. Fusion*, vol. 77, pp. 29–52, Jan. 2022, doi: 10.1016/j.inffus.2021.07.016.
[34] A. R. Javed, W. Ahmed, S. Pandya, P. K. R. Maddikunta, M. Alazab, and T. R. Gadekallu, "A survey of explainable artificial intelligence for smart cities," *Electronics*, vol. 12, no. 4, p. 1020, Feb. 2023, doi: 10.3390/electronics12041020.
[35] L. Butterfield. (2018). *How AI is Shaping the Future of Politics*. [Online]. Available: https://www.research.ox.ac.uk/article/2018-10-15-how-ai-is-shaping-the-future-of-politics
[36] J. Gerlach, P. Hoppe, S. Jagels, L. Licker, and M. H. Breitner, "Decision support for efficient XAI services—A morphological analysis, business model archetypes, and a decision tree," *Electron. Markets*, vol. 32, no. 4, pp. 2139–2158, Dec. 2022, doi: 10.1007/s12525-022-00603-6.
[37] D. G. Broo and J. Schooling, "A framework for using data as an engineering tool for sustainable cyber-physical systems," *IEEE Access*, vol. 9, pp. 22876–22882, 2021, doi: 10.1109/ACCESS.2021.3055652. https://doi.org/10.1109/ACCESS.2021.3055652
[38] M. C. Zizic, M. Mladineo, N. Gjeldum, and L. Celent, "From Industry 4.0 towards Industry 5.0: A review and analysis of paradigm shift for the people, organization and technology," *Energies*, vol. 15, no. 14, p. 5221, Jul. 2022. [Online]. Available: https://www.mdpi.com/1996-1073/15/14/5221
[39] K. Schwab. (2016). *The Fourth Industrial Revolution: What it Means, How to Respond*. [Online]. Available: https://www.weforum.org/agenda/2016/01/the-fourth-industrial-revolution-what-it-means-and-how-to-respond/
[40] I. Mendia, S. Gil-Lopez, I. Grau, and J. D. Ser, "A novel approach for the detection of anomalous energy consumption patterns in industrial cyber-physical systems," *Expert Syst.*, vol. 41, no. 2, p. e12959, Feb. 2024, doi: 10.1111/exsy.12959.
[41] B. A. Y. Alqaralleh, F. Aldhaban, E. A. AlQarallehs, and A. H. Al-Omari, "Optimal machine learning enabled intrusion detection in cyber-physical system environment," *Comput., Mater. Continua*, vol. 72, no. 3, pp. 4691–4707, 2022. [Online]. Available: https://www.techscience.com/cmc/v72n3/47493

[42] A. Procopiou and T. M. Chen, "Explainable AI in machine/deep learning for intrusion detection in intelligent transportation systems for smart cities," in *Explainable Artificial Intelligence for Smart Cities*, 1st ed., M. Lahby, U. Kose, and A. K. Bhoi, Eds. Boca Raton, FL, USA: CRC Press, 2021, ch. 17, pp. 297–321. [Online]. Available: https://www.taylorfrancis.com/chapters/edit/10.1201/9781003172772-17/explainable-ai-machine-deep-learning-intrusion-detection-intelligent-transportation-systems-smart-cities-andria-procopiou-thomas-chen

[43] V. Estrela, O. Saotome, H. Loschi, J. Hemanth, W. Farfan, J. Aroma, C. Saravanan, and E. Grata, "Emergency response cyber-physical framework for landslide avoidance with sustainable electronics," *Technologies*, vol. 6, no. 2, p. 42, Apr. 2018, doi: 10.3390/technologies6020042.

[44] E. Gelenbe and F.-J. Wu, "Future research on cyber-physical emergency management systems," *Future Internet*, vol. 5, no. 3, pp. 336–354, Jun. 2013, doi: 10.3390/fi5030336.

[45] A. K. Luhach and A. Elçi, *Artificial Intelligence Paradigms for Smart Cyber-Physical Systems*. Hershey, PA, USA: IGI Global, 2020.

[46] E. Onyema, C. Edeh, U. Gregory, V. Edmond, A. Charles, and N. Richard-Nnabu, "Cybersecurity awareness among undergraduate students in Enugu Nigeria," *Int. J. Inf. Secur., Privacy Digit. Forensics*, vol. 5, no. 1, pp. 34–42, 2021.

[47] Y. Hu, J. Wu, G. Li, J. Li, and J. Cheng, "Privacy-preserving few-shot traffic detection against advanced persistent threats via federated meta learning," *IEEE Trans. Netw. Sci. Eng.*, vol. 11, no. 3, pp. 2549–2560, May/Jun. 2024. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/10214668

[48] G. Li, J. Wu, S. Li, W. Yang, and C. Li, "Multitentacle federated learning over software-defined industrial Internet of Things against adaptive poisoning attacks," *IEEE Trans. Ind. Informat.*, vol. 19, no. 2, pp. 1260–1269, Feb. 2023, doi: 10.1109/TII.2022.3173996.

[49] G. Li, K. Ota, M. Dong, J. Wu, and J. Li, "DeSVig: Decentralized swift vigilance against adversarial attacks in industrial artificial intelligence systems," *IEEE Trans. Ind. Informat.*, vol. 16, no. 5, pp. 3267–3277, May 2020, doi: 10.1109/TII.2019.2951766. https://doi.org/10.1109/TII.2019.2951766

[50] G. Makridis, S. Theodoropoulos, D. Dardanis, I. Makridis, M. M. Separdani, G. Fatouros, D. Kyriazis, and P. Koulouris, "XAI enhancing cyber defence against adversarial attacks in industrial applications," in *Proc. IEEE 5th Int. Conf. Image Process. Appl. Syst. (IPAS)*, Dec. 2022, pp. 1–8.

[51] N. Nikolakis, V. Maratos, and S. Makris, "A cyber physical system (CPS) approach for safe human–robot collaboration in a shared workplace," *Robot. Comput.-Integr. Manuf.*, vol. 56, pp. 233–243, Apr. 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0736584517302168

[52] A. F. Markus, J. A. Kors, and P. 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," *J. Biomed. Informat.*, vol. 113, Jan. 2021, Art. no. 103655, doi: 10.1016/j.jbi.2020.103655.

[53] S. K. Jagatheesaperumal, Q.-V. Pham, R. Ruby, Z. Yang, C. Xu, and Z. Zhang, "Explainable AI over the Internet of Things (IoT): Overview, state-of-the-art and future directions," *IEEE Open J. Commun. Soc.*, vol. 3, pp. 2106–2136, 2022, doi: 10.1109/OJCOMS.2022.3215676.

[54] G. Srivastava, R. H. Jhaveri, S. Bhattacharya, S. Pandya, Rajeswari, P. K. R. Maddikunta, G. Yenduri, J. G. Hall, M. Alazab, and T. R. Gadekallu, "XAI for cybersecurity: State of the art, challenges, open issues and future directions," 2022, *arXiv:2206.03585*.

[55] Z. M. Aljazzaf, M. Perry, and M. A. M. Capretz, "Online trust: Definition and principles," in *Proc. 5th Int. Multi-Conf. Comput. Global Inf. Technol.*, Sep. 2010, pp. 163–168, doi: 10.1109/ICCGI.2010.17.

[56] K. A. Hoff and M. Bashir, "Trust in automation: Integrating empirical evidence on factors that influence trust," *Human Factors, J. Human Factors Ergonom. Soc.*, vol. 57, no. 3, pp. 407–434, May 2015, doi: 10.1177/0018720814547570.

[57] A. Hoenig and J. D. W. Stephens, "Decision making using automated estimates in the classification of novel stimuli," in *Proc. Int. Conf. Appl. Human Factors Ergonom.*, 2019, pp. 25–35. [Online]. Available: https://api.semanticscholar.org/CorpusID:198313203

[58] M. H. Jarrahi, "Artificial intelligence and the future of work: Human-AI symbiosis in organizational decision making," *Bus. Horizons*, vol. 61, no. 4, pp. 577–586, Jul. 2018, doi: 10.1016/j.bushor.2018.03.007.

[59] R. Eramo, V. Muttillo, L. Berardinelli, H. Bruneliere, A. Gomez, A. Bagnato, A. Sadovykh, and A. Cicchetti, "AIDOaRt: AI-augmented automation for DevOps, a model-based framework for continuous development in cyber-physical systems," in *Proc. 24th Euromicro Conf. Digit. Syst. Design (DSD)*, Sep. 2021, pp. 303–310. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/9556443

[60] T. Khan, K. Ahmad, J. Khan, I. Khan, and N. Ahmad, "An explainable regression framework for predicting remaining useful life of machines," in *Proc. 27th Int. Conf. Autom. Comput. (ICAC)*, Sep. 2022, pp. 1–6.

[61] V. Terziyan and O. Vitko, "Explainable AI for Industry 4.0: Semantic representation of deep learning models," *Proc. Comput. Sci.*, vol. 200, pp. 216–226, Jan. 2022, doi: 10.1016/j.procs.2022.01.220.

[62] Z. A. E. Houda, B. Brik, and L. Khoukhi, "'Why should I trust your IDS?': An explainable deep learning framework for intrusion detection systems in Internet of Things networks," *IEEE Open J. Commun. Soc.*, vol. 3, pp. 1164–1176, 2022.

[63] K. Amarasinghe, "Explainable neural networks based anomaly detection for cyber-physical systems," Ph.D. dissertation, Dept. Comput. Sci., Virginia Commonwealth Univ., Richmond, VA, USA, 2019. [Online]. Available: https://scholarscompass.vcu.edu/etd/6091

[64] S. Pattepu, A. Mukherjee, S. Routray, P. Mukherjee, Y. Qi, and A. Datta, "Multi-antenna relay based cyber-physical systems in smart-healthcare NTNs: An explainable AI approach," *Cluster Comput.*, vol. 26, no. 4, pp. 2259–2269, Aug. 2023, doi: 10.1007/s10586-022-03632-0.

[65] É. Houzé, A. Diaconescu, J.-L. Dessalles, D. Mengay, and M. Schumann, "A decentralized approach to explanatory artificial intelligence for autonomic systems," in *Proc. IEEE Int. Conf. Autonomic Comput. Self-Organizing Syst. Companion (ACSOS-C)*, Aug. 2020, pp. 115–120.

[66] M. Mongelli, "Design of countermeasure to packet falsification in vehicle platooning by explainable artificial intelligence," *Comput. Commun.*, vol. 179, pp. 166–174, Nov. 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0140366421002504

[67] A. Al Hamoud, A. Hoenig, and K. Roy, "Sentence subjectivity analysis of a political and ideological debate dataset using LSTM and BiLSTM with attention and GRU models," *J. King Saud Univ. Comput. Inf. Sci.*, vol. 34, no. 10, pp. 7974–7987, Nov. 2022, doi: 10.1016/j.jksuci.2022.07.014.

[68] D. Sonntag, S. Zillner, S. Chakraborty, A. Lorincz, E. Strommer, and L. Serafini, "The medical cyber-physical systems activity at EIT: A look under the hood," in *Proc. IEEE 27th Int. Symp. Comput.-Based Med. Syst.*, May 2014, pp. 351–356.

[69] J. O. Grady, *System Requirements Analysis*, 1st ed. Amsterdam, The Netherlands: Elsevier, 2006.

[70] U. Schlegel, H. Arnout, M. El-Assady, D. Oelke, and D. A. Keim, "Towards a rigorous evaluation of XAI methods on time series," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshop (ICCVW)*, Oct. 2019, pp. 4197–4201, doi: 10.1109/ICCVW.2019.00516.

[71] É. Houzé, "A generic and adaptive approach to explainable AI in autonomic systems: The case of the smart home," Ph.D. dissertation, Institut Polytechnique de Paris, Palaiseau, France, 2022. [Online]. Available: https://theses.hal.science/tel-03721520

[72] F. Björck, M. Henkel, J. Stirna, and J. Zdravkovic, "Cyber resilience—Fundamentals for a definition," *Adv. Intell. Syst. Comput.*, vol. 353, no. 7, pp. 311–316, 2015.

[73] Z. A. El Houda, B. Brik, and S.-M. Senouci, "A novel IoT-based explainable deep learning framework for intrusion detection systems," *IEEE Internet Things Mag.*, vol. 5, no. 2, pp. 20–23, Jun. 2022.

[74] S. C. Suh, U. J. Tanik, J. N. Carbone, and A. Eroglu, *Applied Cyber-Physical Systems*, 1st ed. New York, NY, USA: Springer, 2014, doi: 10.1007/978-1-4614-7336-7.

[75] J. M. Darias, M. Caro-Martínez, B. Díaz-Agudo, and J. A. Recio-Garcia, "Using case-based reasoning for capturing expert knowledge on explanation methods," in *Proc. Int. Conf. Case-Based Reasoning*. Cham, Switzerland: Springer, 2022, pp. 3–17.

[76] I. Taj and N. Zaman, "Towards industrial Revolution 5.0 and explainable artificial intelligence: Challenges and opportunities," *Int. J. Comput. Digit. Syst.*, vol. 12, no. 1, pp. 285–310, Jul. 2022.

[77] A. B. Bakker, E. Demerouti, and A. I. Sanz-Vergel, "Burnout and work engagement: The JD-R approach," *Annu. Rev. Organizational Psychol. Organizational Behav.*, vol. 1, no. 1, pp. 389–411, Mar. 2014, doi: 10.1146/annurev-orgpsych-031413-091235.

AMBER HOENIG is currently pursuing the Ph.D. degree with the Computational Data Science and Engineering Program, North Carolina Agricultural and Technical State University. She is also a Research Assistant with the Center for Trustworthy AI (CTA), North Carolina Agricultural and Technical State University. Her research interests include explainable artificial intelligence, machine learning, deep learning, natural language processing, data science, big data analysis, the Internet of Things, and cyber-physical systems.

KAUSHIK ROY is currently a Professor and the Chair of the Department of Computer Science, North Carolina Agricultural and Technical State University. His research is funded by the National Science Foundation (NSF), the Department of Defense (DoD), and the Department of Energy (DoE). He is also the Director of the Center for Cyber Defense (CCD) and Center for Trustworthy AI (CTA). He also directs the Cyber Defense and AI (CDA) Laboratory. He has more than 190 publications, including 50 journal articles and a book. His current research interests include cybersecurity, cyber identity, biometrics, machine learning (deep learning), data science, cyber-physical systems, and big data analytics.

YAA TAKYIWAA ACQUAAH received the Bachelor of Science degree in mathematics from the Kwame Nkrumah University of Science and Technology, in 2009, the Master of Philosophy degree in computational nuclear science and engineering from the University of Ghana, in 2013, and the Ph.D. degree in computational science and engineering from North Carolina Agricultural and Technical State University, Greensboro, USA. She is currently a Postdoctoral Scholar with the Center for Trustworthy AI (CTA), North Carolina Agricultural and Technical State University. She has engaged in diverse projects spanning areas such as transportation, object detection, intelligent text data mining, smart sensor networks for occupancy, and machine learning techniques for thermal preference-based HVAC control automation in smart buildings. Her current research interest includes pioneering work in anomaly detection within cyber-physical systems. Leveraging her proficiency in data science methodologies, she actively pursues innovative approaches in her quest for enhancing system reliability and security.

SUN YI received the B.S. degree in mechanical and aerospace engineering from Seoul National University, South Korea, in 2004, and the M.S. and Ph.D. degrees in mechanical engineering from the University of Michigan, Ann Arbor, MI, USA, in 2006 and 2009, respectively. He is currently a Boeing Endowed Professor in mechanical engineering with North Carolina Agricultural and Technical State University. He has developed new and novel methods for analysis and control algorithms of dynamic systems which are reliable and robust. The control methods have been applied to aerospace and autonomous vehicles, networked robots, advanced manufacturing through artificial intelligence, and machine learning. His research has been supported by the DoD, NASA, the Department of Energy, and the Department of Transportation.

SALIL S. DESAI is currently an University Distinguished Professor and the Director of the Center of Excellence in Product Design and Advanced Manufacturing, North Carolina Agricultural and Technical State University. He is also an active researcher in the interdisciplinary fields of smart additive manufacturing, multiscale modeling, product design, and realization with applications in the energy, automotive, aerospace, and biomedical fields. His research interfaces cyber-physical systems, AI algorithms, and computational material design for Industry 5.0. He is a fellow of IISE, ASME, and AIMBE professional societies and an Associate Editor of *IISE Transactions*.