Explainable AI for human-centric ethical IoT systems

Article

Accepted Version

Ambritta P., N. ORCID: https://orcid.org/0000-0001-6310-9378, Mahalle, P. N. ORCID: https://orcid.org/0000-0001-5474-6826, Patil, R. V. ORCID: https://orcid.org/0000-0003-1073-4297, Dey, N. ORCID: https://orcid.org/0000-0001-8437-498X, Crespo, R. G. ORCID: https://orcid.org/0000-0001-5541-6319 and Sherratt, R. S. ORCID: https://orcid.org/0000-0001-7899-4445 (2023) Explainable AI for human-centric ethical IoT systems. IEEE Transactions on Computational Social Systems. ISSN 2329-924X doi: https://doi.org/10.1109/tcss.2023.3330738 Available at https://centaur.reading.ac.uk/114306/

Explainable AI for Human-Centric Ethical IoT Systems

Nancy Ambritta P., Parikshit N. Mahalle, *Senior Member, IEEE*, Rajkumar V. Patil, Nilanjan Dey, *Senior Member, IEEE*, Rubén González Crespo, *Senior Member, IEEE*, and R. Simon Sherratt, *Fellow, IEEE*

***Abstract*—The current era witnesses the notable transition of the society from an information centric to a human-centric one aiming at striking a balance between economical advancements and upholding societal and fundamental needs of humanity. It is undeniable that Internet of Things (IoT) and Artificial Intelligence (AI) are the key players in realizing the human-centric society. However, for society, and individuals to benefit from the advanced technology, it important to gain the trust of the human users by guaranteeing the inclusion of ethical aspects such as safety, privacy, non-discrimination, and legality of the system. Incorporating Explainable AI (XAI) into the system to establish explainability and transparency supports the development of trust amongst the stakeholders including the developers of the system. This paper proposes a framework for a human centric IoT system with Explainable AI that provides explanations for a particular decision by the AI model. Further, we incorporate mechanisms to improve the system from providing mere explanations to decisions, into systems that are interpretable, context adequate, and actionable. The XAI framework will consider all possible future events with quantifiable values assigned to features and outcomes, enabling the users to undertake well informed decisions.**

***Index Terms*—Explainable AI (XAI), Ethical AI, Trustworthy AI, privacy, security, human-centric AI, contextual AI, actionable AI, interpretability, Society 5.0.**

I. Introduction

THE Internet of Things (IoT) across various application domains uses Artificial Intelligence (AI) techniques for various activities, including automated decision making, learning from user actions, tracking, tracing user activity, and event notification. While most of the AI techniques are promising in terms of performance and accuracy, the understandability of these models and their outcomes is a recent point of concern with the new laws being introduced that uphold the fundamental rights of the users. Applications in non-critical domains such as agriculture, hospitality etc., use AI algorithms to learn from sensor inputs and propose actions directed towards attaining the target features. In such cases, the human users depend upon the system's decisions to carry out tasks without understanding the reason behind the decision or outcome. The users eventually begin to trust the system with a tolerance to the minimal error percentage that may occur. However, in critical application domains such as healthcare, traffic safety etc., a collaboration between the machine and the human users is required. Inclusion of human users in the system mandates addressing important ethical aspects pertaining to the fundamental rights of humans. Albeit a lot of debate exists on the ways to express ethical AI. The question is whether it can be called AI ethics, responsible AI, or trustworthy AI. Trustworthy AI may fit more appropriately. We want people to understand and trust the AI technology that it is well built, and it does what it is meant to do.

Globally, there is not one way to look at and explain ethics, as different cultures, societies, and beliefs come into existence. However, everybody understands what trust is and how we build that will be the crux of the ethics that we deal with. Understanding the stakeholders and government structure in key regions is important in developing the trustworthy AI. The four pillars that are the primary elements for a trustworthy AI are privacy, safety & security, non-discrimination & elimination of bias, explainability, and transparency [1]. The need for the four pillars is because, if we are speaking about trust then we need to comply with certain regulations. Although a lot of research has been undertaken in the fields of privacy and safety and security, the two pillars that are newer and need understanding are the fields of non-discrimination and bias, and explainability. Explainability involves two aspects. First, is to whom we are trying to explain and the next is what we are trying to explain. Our focus in this research is to address the explainability and interpretability issues in human centric IoT systems in addition to the privacy and security aspects. This requires the AI system to present the reasons towards taking up a particular course of action to the humans in an understandable form, thereby upholding the ethical values. Fig 1 shows a typical IoT system with AI framework in a human-centric society. The IoT sensors that collect data from the surroundings,

Submitted September 2022. *(Corresponding author: Nancy Ambritta P.).*

N. Ambritta P. is with Glareal Software Solutions PTE. LTD, Singapore 200809 (e-mail: nancy.ambritta@glareal.com).

P. N. Mahalle is with the Department of Artificial Intelligence and Data Science, Vishwakarma Institute of Information Technology, Pune 411048, India (e-mail: parikshit.mahalle@viit.ac.in).

R V. Patil is with the Department of Information Technology, Vishwakarma Institute of Information Technology, Pune 411048, India (e-mail: rajkumar.patil@viit.ac.in).

N. Dey is with Techno International New Town, Kolkata, 700156, India (e-mail: nilanjan.dey@tint.edu.in).

Rubén González Crespo is with the Department of Computer Science and Technology, Universidad Internacional de La Rioja, Logroño, Spain (e-mail: ruben.gonzalez@unir.net).

R. Simon Sherratt is with the Department of Biomedical Engineering, the University of Reading, Reading, UK (e-mail: r.s.sherratt@reading.ac.uk).

Color versions of one or more of the figures in this article are available online at http://ieeexplore.ieee.org

and the software applications that are built for a particular IoT application to facilitate interaction and operation shown in Fig. 1 are typical in any IoT application. They are connected over the Internet over which resourceful operations such as AI processing of data and analysis, storage, computation, and communication occur. Human users interact with each of these layers based upon the level of access, purpose, and usage such as the developers, end users, system admins and the technical support team as presented in Fig. 1.

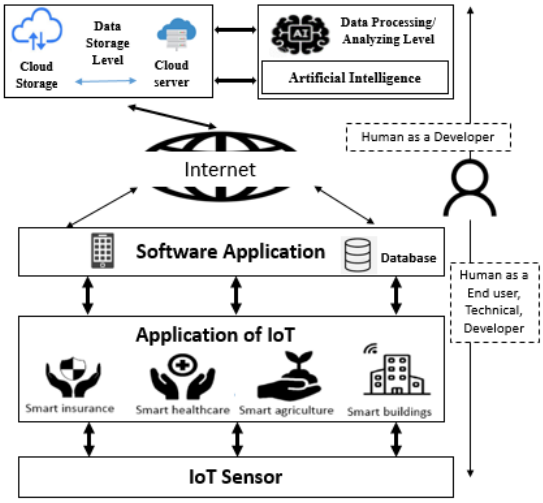

Fig. 1. A typical IoT system with AI framework in a human-centric society.

This paper proposes a human inclusive Explainable AI (XAI) framework for IoT that supports informed and transparent decision making to the end users in critical applications. The framework proposes to consider all possible outcomes upon the choice of a particular course of action thereby making the system actionable. The changes or improvement in outcomes can be assessed by the end users by toggling between selected features involved in the decision-making process thereby making the system interpretable. Furthermore, solutions to presenting explanations pertaining compliance to privacy and security aspects have been addressed in the proposed framework. An attack model based on the study of possible attacks to the system is also presented in the paper.

The rest of the paper is organized as the follows. We present and discuss the challenges in realizing a trustworthy human centric ethical IoT system in section II. Here we highlight the various possible security attacks possible in an IoT system followed by presenting an attack model based on the discussion. Various mitigation techniques are discussed which helps in identifying a suitable technique for our proposed system. In this view emphasis has been made on the intrusion detection mechanisms from existing literature. In section III we discuss the application of XAI in various autonomous decision making IoT applications that emphasize the inclusion of human users. The literature presented in this section elaborates on use of XAI in providing explanations to various intrusions detected in the system. In section IV we discuss the various factors that make explainability in a human centric IoT system to be a difficult yet not an impossible task. This will shed light on the aspects that must be considered while introducing explainability into the IoT system. In section V we present our proposed architecture that uses two XAI frameworks, one to explain the detected intrusion in the system which will also include features of actionability and interpretability, and the other to explain the incorporation of data protection rules in the system which will improve the trustworthiness of the system amongst the stakeholders. Section VI presents our conclusions and potential further work.

II. Challenges in Trustworthy Human-Centric Ethical IoT Systems

The major ethical aspect to be considered in Society 5.0, human in the loop model, is to ensure that the system is trustworthy. Trustworthiness of a system depends upon the system's ability to ensure confidentiality, integrity, availability, and accountability. Although, various systems claim to ensure all the above criteria, it is important to involve the stakeholders in the process and make the system more transparent and interpretable. This section discusses the various cyber threats in the cyber physical systems that are also possible in medical cyber physical systems, leading to disruption in services and malfunctioning. Attacks to the system happen by exploiting the vulnerabilities in the system that exists in the two spaces (cyber space and physical space) of the IoT system. Attackers either breach the privacy or try to directly influence the system and compromise the functioning. The common sources or entry points for the attackers are through communication links, software, platform/hardware, and users in general.

A. Vulnerabilities

Vulnerabilities in IoT systems occur due to the following reasons, namely, weak identity management and access control, lack of security policies by design, platform vulnerabilities, network, or communication link vulnerabilities. The weak identity and access management leads to many insider attacks in the system who may leak sensitive information to unauthorized users and access/tamper sensitive devices causing damage to the end users. Insider attacks affect all the security aspects such as confidentiality, integrity, and availability [2]. The confidentiality of the system is compromised due to weak encryption of data in transit thereby enabling attacks to eavesdrop over the communication link and tamper/steal information [2]. Vulnerabilities in the software enable the attacks to affect the integrity of the system by performing memory related attacks such as buffer overflow attack and malware injection to modify the critical decision variables and sensor information [3]. False data packet injection and replay attacks affect the integrity of the system by modifying the information or injecting new fabricated information [4]. The physical layer/platform/hardware is also prone to attacks such as sensor data spoofing and tampering of configurations affecting the security of the system [5]. DoS attacks are major contributors affecting the availability of the system. Attacks

such as traffic jamming affects normal and relevant communication by creating interference and collision of requests over the communication link. Broadcasting spoofed network packets, causing an overflow of the buffer memory and corruption of data or modification of physical configurations of devices also affect the availability of the system. Other vulnerabilities in the system include weak personal data management, increased heterogeneity and connectivity and lack of security policies by design. With the above discussed vulnerabilities, we can now list down the possible attacks on the system and further model the attacks. Table I presents various vulnerabilities and the associates attacks that affect the system's confidentiality, integrity, availability, and accountability of the system. A pictorial representation of the various vulnerabilities and attacks that affect the confidentiality, integrity, and availability of the IoT system is also shown in Fig. 2.

B. Attack Modelling

Fig. 3 shows the attack model in a human centric IoT system. The attacks labelled 1-7 in Fig. 3 are discussed below with reference to the attacks discussed in Fig 2.

TABLE I
VULNERABILITIES AND ATTACKS IN AN IOT SYSTEM

Vulnerabilities	Attacks		
	Confidentiality	Integrity	Availability
Weak identity and access control	1) Insider Leak 2) Inadvertent leak	7) Insider tampering	11) Insider manipulation
Platform	3) Hardware hacking	8) Sensor spoofing	12) Configuration modification
Network and communication link	4) Eavesdropping	9) Packet injection 10) Replay attack	13) Jamming 14) Flooding
Software	5) Malware 6) Buffer overflow	Malware, Buffer overflow	Malware, Buffer overflow

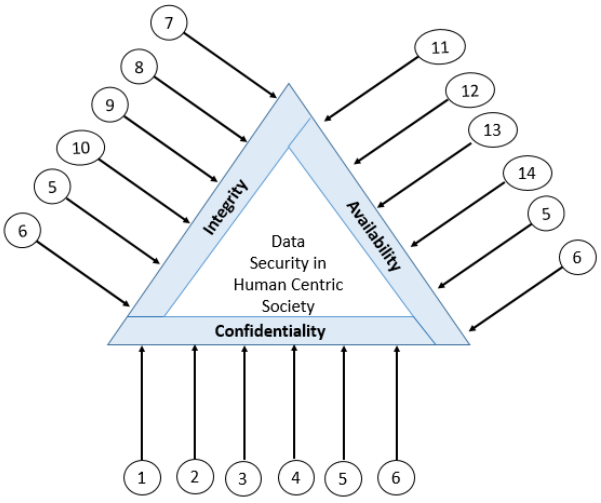

Fig. 2. Vulnerabilities and Attacks in the IoT system.

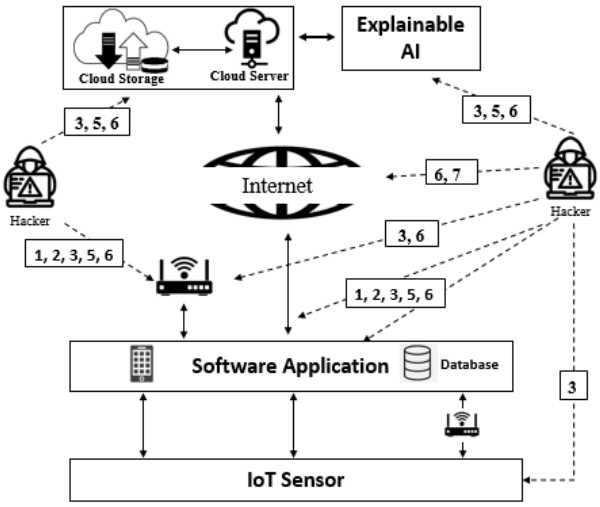

Fig. 3. Attack modelling.

1) Spoofing

Spoofing refers to the attack where a suspicious agent portrays to be a legitimate entity by stealing the identity of a legitimate user and attempts to gain sensitive information from users with an intent to cause damage. Spoofing occurs in a variety of forms and a few of them are sensor spoofing, packet injection, eavesdropping and replay attacks that also affects the integrity of the system.

2) Phishing

Phishing is an attack on the system wherein an attacker lures the users into sharing sensitive information by means of social engineering. Malware, packet injection and replay attacks are forms of phishing attacks on the system.

3) Tampering

Tampering is an attack on the system that targets application parameters, manipulates data or information in transit thereby affecting the integrity of the system. Attacks such as insider tampering, insider manipulation, packet injection, replay attack and malwares are a few forms of tampering attacks.

4) Repudiation

Repudiation attacks are performed with the goal of making a suspicious activity untraceable or claiming that a communication or exchange of data never happened. Improper tracking or log maintenance, poor identity and access control techniques leads to repudiation attacks such as the replay attack, packet injection, insider tampering, insider manipulation, configuration manipulation and hardware hacking.

5) Information disclosure

Information disclosure generally affects the confidentiality of the system. It is the inability of an application to secure sensitive data that is not meant to be exposed to users without proper access. Exposure of vulnerabilities in the system is also a form of information disclosure that would enable the attacks to exploit and gain access to the system. Insider leak, Inadvertent leak, hardware hacking, eavesdropping, malware and buffer overflow are a few attacks that contribute to information disclosure.

6) Denial of Service (DoS)

DoS attacks are performed to hamper the normal activities of the system thereby affecting its availability. Jamming, flooding, malware, and buffer overflow are few forms of the DoS attacks.

7) Privilege Elevation

Privilege Elevation leads to unintended attacks on the system by providing unauthorized access to external as well as internal users paving way for attacks such as insider leak, insider tampering, insider manipulation, hardware hacking and platform configuration manipulation. To counter this proper identity and access management schemes should be in place.

C. Mitigation Techniques

The known methods in securing the medical cyber physical systems fall under the three broad categories namely, the intrusion detection mechanism, cryptographic methods, and system hardening.

Cryptographic measures refer to the usage of encryption techniques to protect the communication channels and the messages in transit from unintended access and manipulation. However, the traditional cryptographic measures are not suitable for the human centric IoT system as the computations require high energy and resources which are limited in the IoT devices. The problem of overhead has been addressed through alternate approaches such as compression techniques used before applying encryption and light weight cryptographic techniques as proposed by Masud *et al.* [6]. Similarly, Ullah *et al.* [7] proposed a lightweight scheme to ensure confidentiality and authentication by combining the encryption and digital signature methods in a single step process. Shamshad *et al.* [8], proposed a lightweight key establishment protocol to secure a patient's physiological datum. Hahn *et al.* [9] proposed an effective countermeasure for securing the resource-limited mobile healthcare systems while still providing effective access control and delegation mechanisms. Xu *et al.* [10] proposed a blockchain based trustworthy edge caching scheme to improve the Quality of experience for mobile users.

System hardening refers to the measures taken to combat the attacks that occur in the IoT system due to interfacing with external devices, software, and execution environments. By utilizing the hardware security models such as Intel's TrustLite [11] and TrustZone ARM, applications that are very critical and require more security can be isolated from untrusted OS. Liu *et al.* [12] proposed a trust detection based secured routing (TDSR) scheme that provides a secured route to carry data packets from source nodes to data centers. The problem that arises due to the heterogeneity of the connected devices from various networks can be handled by inter-authentication of devices. Renuka *et al.* [13] proposed a secure password-based authentication scheme that facilitates participating entities in a M2M network to mutually authenticate each other to share data securely with the help of a shared session key. Shepherd *et al.* [14] presented an analysis of various trusted computing technologies in the CPS domain such as the Trusted Platform Module (TPM), Secure Elements (SE), Hypervisors and Virtualization, Intel TXT, Trusted Execution Environments (TEE) and Encrypted Execution Environments.

Intrusion detection includes monitoring the IoT system at runtime for any suspicious activity. The intrusion detection mechanisms generally fall under two broad categories namely the data centric approaches and specification centric approaches. In the data centric approach, the measurements collected from the physical devices help in detection of intrusions. The anomaly-based intrusion detection and device/network fingerprint mechanisms are examples of the data centric detection approaches can be seen in Keshk *et al.* [15] and Zhou *et al.* [16]. Although the data centric methods make it easy to apply machine learning techniques to the cyber and physical domains irrespective of the characteristics of the CPS, neglecting the domain specific details of the CPS is critical in analyzing the impact of the attacks.

The specification centric approach makes use of already established standards, rules, and specifications of the specific CPS models to detect anomalies and inconsistencies. However, designing specification centric detection mechanisms is often a challenging task, since revealing the design documents to third party security monitoring bodies could affect the business and incur loss. Mitchell and Chen [17] proposed a mechanism to transform the behavior of a physical device being monitored into a state machine against which the behavior can be verified to detect any deviations. Mowla *et al.* [18] proposed a lightweight classification algorithm that runs locally on the device, thereby preserving the privacy of the data and reducing the overhead on the communication channel. A hybrid detection method that includes the advantages of both the data centric and specification centric detection mechanisms are widely used across the CPS domains. Fauri *et al.* [19] demonstrated that the combined hybrid method of leveraging the measurements/characteristics collected from the cyber domain for anomaly detection paired with the mathematical models of the physical devices (state machines etc.), can effectively detect the attacks to the system. As discussed earlier, most of the attacks to an IoT system occur in the cyber domain that cause disruptions to the physical layer leading to physical device malfunction, threat to safety of patients, financial loss and hampering the observability of the physical devices [20]. For an IoT system, the aspect/feature that is critical in designing the detection methods is the cyber and physical domain characteristics of the involved robots and devices. Timely detection of attacks in the cyber layer is very important before the effects are applied to the physical devices. The hybrid detection approach is found to be handy in such cases wherein a dynamic behavioral model of the physical devices/ actuators to measure the effects of the control commands before the actual application and an anomaly detection module for continued monitoring of measurements in the cyber domain [21].

III. Using XAI to Explain Intrusions

While many machine learning and deep learning models help in effectively detecting intrusion in cyber physical systems, the transparency and interpretability of the models is still a question to be addressed considering the emerging human-in-the-loop society,

i.e., society 5.0. Involving the stakeholders in the process via XAI provides avenue for improvement of the model and gaining trust amongst the users. Table II shows a summary of the literature that uses XAI methodologies to address the security and efficiency problems in various domains. Zebin *et al.* [22] addressed the issue that exists in the DNS over HTTPS (DoH) protocol wherein the network administrators are incapable of identifying malicious network traffic generated by malware and suspicious tools. The authors have developed a novel machine learning framework using the random forest algorithm to identify and classify the DNS over HTTPS attacks. Also, XAI has been utilized to highlight the features that contributed towards the classification thereby making the results transparent and explainable. Hwang and Lee [23] proposed a mechanism using XAI to visually display the sensors that are behaving abnormally when an intrusion happens to reduce the overhead of multiple checks in the event of a false alarm. Moraliyage *et al.* [24] proposed a novel multimodal classification approach for deep learning algorithms, that enable the identification and classification of the onion services in the dark web. The anonymity of the services and the complexity of the Tor's HS protocol upon which the dark web operates makes it difficult for the cyber threat intelligence software to identify these services with criminal intent. Here, the authors have used XAI technology to classify and contextualize the features of these onion services. Suryotrisongko *et al.* [25] proposed a mechanism to identify and classify the DGA-based botnets based upon statistical features, using the random forest AI algorithm. XAI technologies along with the Open-source intelligence (OSINT) has been utilized to improve the model's explain ability of the output thereby improving the trustworthiness of the model. Li *et al.* [26] proposed a mechanism for the detection of Advanced Persistent Threats (APT) at the resource constrained edge devices using XAI and CTI. The defense mechanisms and resource allocation at the edge devices is designed and governed by the explanations and guidelines provided by the proposed mechanism. Mane and Rao [27] demonstrated the intrusion detection mechanism on the NSL KDD Dataset that contains normal and attack data. A Deep neural network has been used to detect the attack and XAI has been applied to explain the prediction based upon features at every stage of the machine learning pipeline. Table II shows a consolidated view of the various applications of XAI in the existing literature.

IV. Explainability issues in Human Centric IoT Systems

Although XAI has been effectively used in a wide range of domains, the automated decision-making process in critical application domains, such as banking, healthcare etc., raise concerns in upholding the fundamental rights of users.

A. Legal Requirements: Expanding XAI to Interpretable and Actionable AI

The legal requirement in critical automated decision-making process emphasizes the involvement of human users. Enabling them to assess the decision process, express their point of view and contest the decision if required is an important provision mention in GDPR. Ensuring the compliance to the data protection rules and regulation laid down by specific government organizations is of prime importance in critical automated decision-making applications. Hamon *et al.* [28] explained the necessity to incorporate explainability in the system to explain the compliance to rules and regulations thereby implementing trustworthiness in the critical automated decision-making systems by design. Furthermore, they emphasized the fact that although mechanisms to document and audit the logic of the underlying algorithm involved in decision making are in place, the increased complexity of the AI based algorithms makes it difficult to present the outcomes in an understandable format to the humans. Upholding the 'right to explanation' is a tedious aspect to address as the evaluation of the relevance of the explanations from a legal perspective and the establishment of strong causal links between the input data and the outcomes is not agreeably established. Understanding the context of the application should also be considered while evaluating the relevance and adequacy of explanations.

While explainability refers to providing an explanation of the system's internal working to the users, interpretability refers to the transition that occurs when the cause and effect of the AI system's decision is understandable to the users. The decision of an AI system should be contextual. Objectives and situations keep varying in real-time. Therefore, it is important that the decisions consider all possible future effects, or the AI systems proposes a decision that considers the most probable future event and presents it to the humans using XAI. This will enable the users to make informed decisions, in critical situations. Further, actionability of the AI systems includes providing a level of confidence associated with a particular course of action. Albeit the incorporation of these features into explainable AI framework comes with a set of challenges as discussed below.

B. Challenges in Providing Human Understandable Explanation for AI-Based Decision-Making Systems

Consider a scenario wherein a person suspects they have a COVID-19 infection and therefore presents themselves at the emergency ward for observation. After a blood test, a nurse conveys reports to suggest a possibility of COVID-19 infection. In such cases, the patient is examined by the doctor and admitted to the intensive care unit anticipating lung damage due to pneumonia. However, when a doctor is not available, an AI based automated decision-making system can make recommendation based on the X-ray images. Bringing the automated decision-making system into the process mandates the implementation of fundamental rights of the users as mentioned in the previous paragraphs. Involving the humans in the process requires explanations. The wide range of technical aspects that challenge the feasibility of providing explanations to AI based models are presented below in this section with the help of the above use case.

1) Complexity of data

The increased storage facilities of devices and digitalization of equipment has supported the collection of diverse data including image, text, tabular data, graphs and many more. The technological assistance in these devices also facilitates the detailed collection of data. For example, in the scenario presented above, the X-ray in medical imaging data consists of

TABLE II
XAI APPLICATIONS IN THE EXISTING LITERATURE

References	Purpose of XAI Research	Detection Mechanism	Domain	XAI Explanation Type	XAI Model	AI Algorithm
Zebin *et al.* [22]	DNS over HTT PS attack	Data centric- anomaly detection (cyber space)	Networking	Feature importance and visualization	SHAP	Random Forest
Hwang and Lee [23]	To reduce false positive ("ESFD")	Data centric- anomaly detection (cyber space)	Industrial Anomaly	Feature Importance	SHAP	Bi-LSTM
Moraliyage *et al.* [24]	Onion services – information trafficking	Data centric- anomaly detection (cyber space)	Dark web-onion services	Visualization	Grad-CAM & attention maps	Bi-LSTM
Suryotrisongko *et al.* [25]	Domain Generation Algorithm based Botnet – Charbot, DeepDGA, MaskDGA	Data centric- anomaly detection (cyber space)	Networking	Feature Importance, summary plot	LIME, SHAP, counterfactuals and ANCHORS	Random Forest
Li *et al.* [26]	Advanced Persistent threats	Data centric- anomaly detection (Physical Edge devices- Device Layer)	Networking (Cloud -Edge devices)	Datapoints	LIME	LSTM & Conv LSTM
Mane and Rao [27]	DoS, Probe, U2R, R2L	Data centric- anomaly detection (cyber space)	Networking	Feature Importance, summary plot	SHAP, LIME, CEM, Protodash, BRCG	Deep Learning Neural Networks

numerable pixels with larger spatial information of organs including various color codes.

2) Complexity of models

The models used in machine learning play an important role in transforming the input data into predictions. The model's complexity is increased by stacking together multiple layers of simple operations to solve complex tasks. This eventually affects the interpretability of the model. For example, in the use case presented above, the deep learning model generally consists of multiple layers with a series of operations and parameters. The deeper layers have complex patterns that are difficult to be interpreted by the practitioners themselves.

3) Complexity of AI algorithms

The development of an AI based system involves a systematic and sequence of steps namely, data processing, training, and evaluation. Implementing these steps involves several processes such as cleaning, data acquisition, feature extraction, sample generation, optimization schemes and many more. After the model has been trained, its performance is evaluated against suitable metrics. The complexity that comes with the incorporation of all the above-mentioned steps makes it difficult to reverse engineer the results/predictions, thereby making it difficult to perform audits on the respective algorithms.

4) Complexity of explanatory techniques

The techniques used for explanations vary depending upon the AI models, since different models depend upon different features for classification and decision making. Hence, in case of medical imaging use case, methods such as occlusion maps. Here the abnormality in a particular region is identified with the help of a prediction score set for a masked region on the image. A high score indicates a non-infection in the masked are. Other methods include gradient descent, counterfactuals etc., all of which do not guarantee that the indicated regions are the ones considered in the decision making. Selection of proper parameters such as the appropriate size of the masked area and a step size for movement also influences the outcome.

5) Trade-off between accuracy and explainability

In general, the two desirable properties of a system include its accuracy (perform computations with less errors) and interpretability (ability to explain the internal workings of the system). However, achieving one property comes at the expense of the other. A method that is interpretable, would involve constraints that reduce the complexity of the system such as reduction in the number of features/parameters to be considered, thereby reducing the accuracy of the model. For example, in the COVID-19 use case, the deep learning methods have increased their accuracy by increasing the complexity of the models. This has made it difficult to provide explanations to such complex systems.

V. PROPOSED FRAMEWORK

In this section we propose a framework that incorporates XAI to provide interpretations and explanations to the stakeholders and present them as actionable outcomes. In our previous work, we have proposed the inclusion of a security layer with security protocols such as OAuth and UMA to implement personal data management following the data protection laws of the country in the design phase itself [29]. Here, we further extend the proposed system by introducing an intrusion detection mechanism/AI model into the system. Further, we incorporate the XAI model that analyzes the AI model's output based on various inputs. The explanation interface provides explanations to the stakeholders in understandable formats such as visualizations, graphs, and reports. A high-level design of the proposed framework is shown in Fig. 4, is an extension to Fig. 1 which shows a typical IoT system with AI framework in a human-centric society. Here, an additional

component, i.e., XAI block has been included that interacts with the AI model, processing and storage components over the cloud/Internet and takes care of the explanations, to provide a reasoning and develop trust with the stakeholders. In our detailed proposed framework shown in Fig. 5. We have incorporated two separate XAI interfaces to address the context adequacy requirement. The first interface takes inputs that are specific to intrusion detection such as packet flow characteristics, network traffic, access control and firewall breaches etc., from the IoT application, communication network and the intrusion detection AI model. It then presents the explanations in appropriate formats such as graphs and data plots that are understandable to the developers and users. The second XAI interface houses the personal data evaluation framework. We present here that the XAI models themselves as a singular tool cannot provide satisfactory explanations to the stakeholders. Regulators are not necessarily technical people and hence we propose to present explanations in the form of reports with supporting tools. The inputs to this interface are event logs and class diagrams that provide important information required for the general data protection laws such as the fundamental data entities in the system, personal data used in the system, data storage location, mechanism and circumstances of data collection, usage of the data, access control and privileges, consent of data collection and usage, security level of the data, path taken for data exchange, and third parties involved in data exchange. The proposed system thereby ensures achieving trustworthiness by expanding 'explainability' into a system that is interpretable and actionable. To aid the interpretability of system in users, they are provided with the facility to toggle between the important features contributing to the decision and check the AI system's reaction to the changes. Further, actionability is implemented by assigning quantifiable values to the various

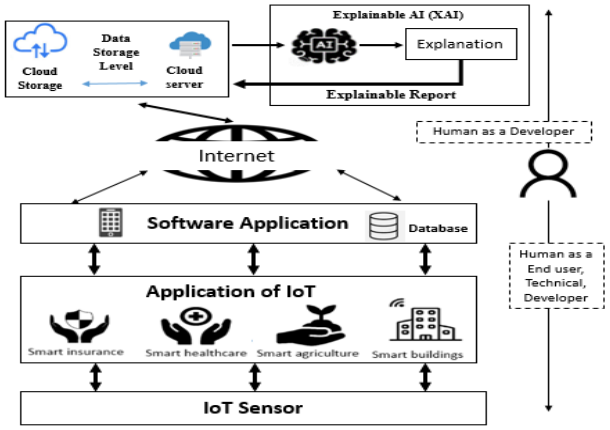

Fig. 4. High level IoT system with Explainable AI framework in a human centric society.

alternatives course of actions provided by the system to improve the trust level in the users. Based upon the values a user can decide on a suitable action plan that will meet the user's needs based on the current situation. The user groups to whom the explanations are presented are general end users of the application, developers, and regulators. The end users use the explanations to choose between alternatives and make critical decisions. The developers use the explanations to understand the finer details of the system and refine the system design. Regulators use the explanations to evaluate the data protection laws and suggest design changes if necessary.

Furthermore, an explanation evaluation framework has been included in the system following the guidelines laid down by DARPA [30]. The evaluation framework also aims at enhancing the trust and appropriate use of the system thereby supporting the human-machine collaboration. The evaluation measures are of

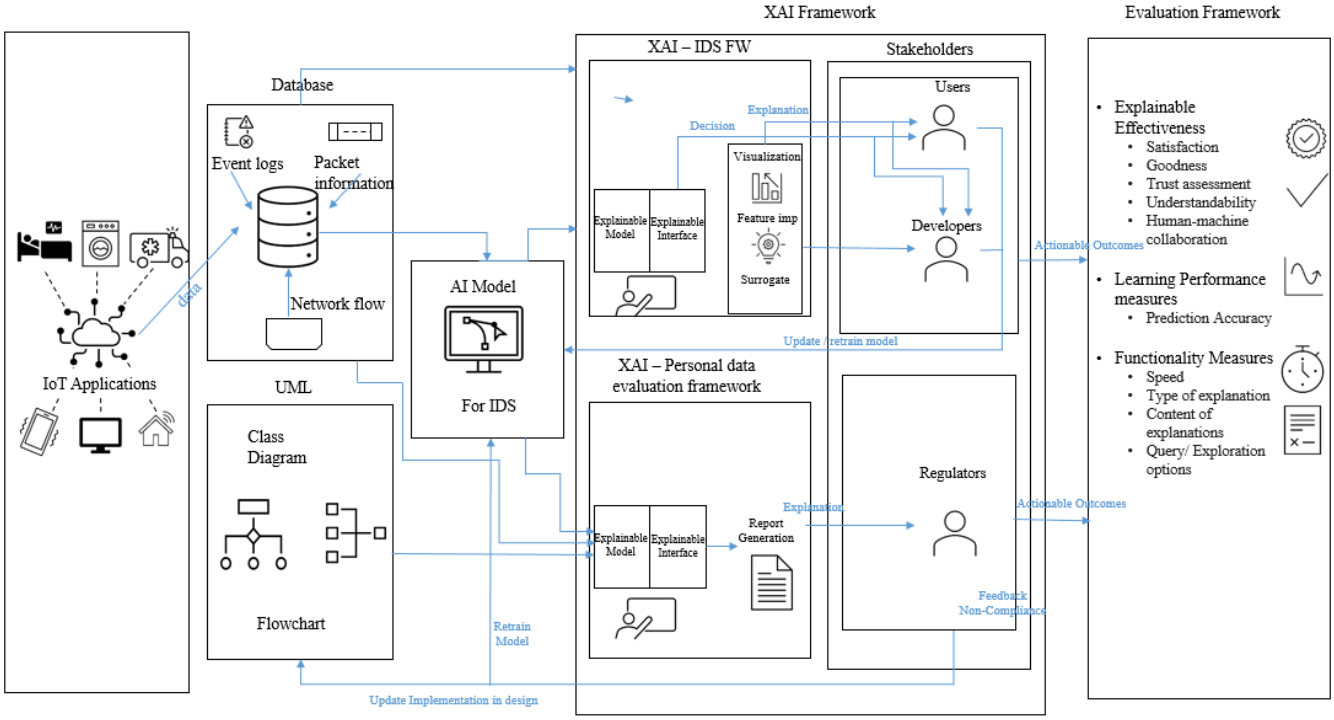

Fig. 5. Proposed Framework for an IoT system with Explainable AI in a human centric society.

three categories, namely, functionality measures, learning performance measures and explainability effectiveness measures. Functionality measures pertain to the speed of generation of explanations, the content (cause, effects, examples, relations etc.), exploration options for users (query, multiple choice etc.) and mode of explanation (visual, text etc.). The learning performance measures represent the accuracy of prediction of the machine learning model. An explanation effectiveness measure refers to the explanations' level of satisfaction, goodness and understanding by the users. The explanation effectiveness can be measured by conducting surveys and interactions with the users and are critical evaluation metrics in XAI.

VI. Conclusions and Future Work

In this paper we have presented an overview of the current IoT system and the ethical aspects surrounding the development an IoT system with human in the loop considering the future society 5.0. The importance of an ethical system for the trustworthiness and upholding of fundamental human rights has been discussed. We have also presented the vulnerabilities and mitigation techniques in the IoT system with supporting literature. The challenges in providing explanations to gain human trust in human centric IoT systems has been discussed by means of a medical imaging use case. Finally, the proposed system has been elaborated in detail which houses the components to realize a context aware, interpretable, actionable, and evaluable explainable AI system that aims at enhancing the usability and trustworthiness in the IoT system. Our future work would be to concentrate on the data collection mechanisms for the IoT system and elimination of bias in data collection that affects the system's performance and decisions in critical tasks.

References

[1] European Commission, "Building trust in human-centric AI," 2018 [Online]. Available: https://ec.europa.eu/futurium/en/ai-alliance-consultation.1.html

[2] L. Cheng, F. Liu, and D. Yao, "Enterprise data breach: causes, challenges, prevention, and future directions," *WIREs Data Mining Knowledge and Discovery*, vol. 7, e1211, Jun. 2017, doi: 10.1002/widm.1211

[3] S. Hanna, R. Rolles, A. Molina-Markham, P. Poosankam, K. Fu, and D. Song, "Take two software updates and see me in the morning: The case for software security evaluations of medical devices," in *Proc. HealthSec '11*, San Francisco, CA, 2011. [Online]. Available: https://spqrlab1.github.io/papers/hanna-aed-healthsec11.pdf

[4] A. Y. Javaid, W. Sun, V. K. Devabhaktuni, and M. Alam, "Cyber security threat analysis and modeling of an unmanned aerial vehicle system," in *Proc. THS*, Waltham, MA, 2013, doi: 10.1109/THS.2012.6459914

[5] T. Trippel, O. Weisse, W. Xu, P. Honeyman, and K. Fu, "WALNUT: Waging doubt on the integrity of MEMS accelerometers with acoustic injection attacks," in *Proc. EuroSP*, Paris, France, 2017, doi: 10.1109/EuroSP.2017.42

[6] M. Masud, G. S. Gaba, S. Alqahtani, G. Muhammad, B. B. Gupta, P. Kumar, and Ahmed Ghoneim, "A lightweight and robust secure key establishment protocol for internet of medical things in COVID-19 patients care," *IEEE Internet Things J.*, vol. 8, no. 21, pp. 15694–15703, Dec. 2020, doi: 10.1109/JIOT.2020.3047662

[7] I. Ullah, N. U. Amin, A. Almogren, M. A. Khan, M. I. Uddin, and Q. Hua, "A lightweight and secured certificate-based proxy signcryption (CB-PS) scheme for e-prescription systems," *IEEE Access*, vol. 8, pp. 199197–199212, Oct. 2020, doi: 10.1109/ACCESS.2020.3033758

[8] S. Shamshad, K. Mahmood, S. Hussain, S. Garg, A. K. Das, N. Kumar, and J. J. P. C. Rodrigues, "An efficient privacy-preserving authenticated key establishment protocol for health monitoring in industrial cyber–physical systems," *IEEE Internet Things J.*, vol. 9, no. 7, pp. 5142–5149, Aug. 2021, doi: 10.1109/JIOT.2021.3108668

[9] C. Hahn, H. Kwon, and J. Hur, "Trustworthy delegation toward securing mobile healthcare cyber-physical systems," *IEEE Internet Things J.*, vol. 6, no. 4, pp. 6301–6309, Aug. 2019, doi: 10.1109/JIOT.2018.2878216

[10] Q. Xu, Z. Su, and Q. Yang, "Blockchain-based trustworthy edge caching scheme for mobile cyber-physical system," *IEEE Internet Things J.*, vol. 7, no. 2, pp. 1098–1110, Feb. 2020, doi: 10.1109/JIOT.2019.2951007

[11] P. Koeberl, S. Schulz, A.-R. Sadeghi, and V. Varadharajan, "TrustLite: a security architecture for tiny embedded devices," in *Proc. EuroSys*, Amsterdam, The Netherlands, 2014, doi: 10.1145/2592798.2592824

[12] Yuxin Liu, Xiao Liu, Anfeng Liu, Neal N. Xiong, and Fang Liu, "A trust computing-based security routing scheme for cyber physical systems," *ACM Trans. Intelligent Systems and Technology*, vol. 10, no. 6, 61, Nov. 2019, doi: 10.1145/3321694

[13] K. M. Renuka, S. Kumari, D. Zhao, and L. Li, "Design of a secure password-based authentication scheme for M2M networks in IoT enabled cyber-physical systems," *IEEE Access*, vol. 7, no. 5, pp. 51014-51027, Mar. 2019, doi: 10.1109/ACCESS.2019.2908499

[14] C. Shepherd, G. Arfaoui, I. Gurulian, R. P. Lee, K. Markantonakis, R. N. Akram, D. Sauveron, and Emmanuel Conchon, "Secure and trusted execution: Past, present, and future - a critical review in the context of the internet of things and cyber-physical systems," in *Proc. TrustCom*, Tianjin, China, 2016, doi: 10.1109/TrustCom.2016.0060

[15] M. Keshk, E. Sitnikova, N. Moustafa, J. Hu, and I. Khalil, "An integrated framework for privacy-preserving based anomaly detection for cyber-physical systems," *IEEE Trans. Sustain. Comput.*, vol. 6, no. 1, pp. 66–79, Jan.-Mar. 2021, doi: 10.1109/TSUSC.2019.2906657

[16] X. Zhou, W. Liang, S. Shimizu, J. Ma, and Q. Jin, "Siamese neural network based few-shot learning for anomaly detection in industrial cyber-physical systems," *IEEE Trans. Ind. Informat.*, vol. 17, no. 8, pp. 5790–5798, Aug. 2021, doi: 10.1109/TII.2020.3047675

[17] R. Mitchell and I. Chen, "Behavior rule specification-based intrusion detection for safety critical medical cyber physical systems," *IEEE Trans. Dependable Secure Comput.*, vol. 12, no. 1, pp. 16–30, Jan.-Feb. 2025, doi: 10.1109/TDSC.2014.2312327

[18] N. I. Mowla, I. Doh, and K. Chae, "On-device AI-based cognitive detection of bio-modality spoofing in medical cyber physical system," *IEEE Access*, vol. 7, pp. 2126–2137, Dec. 2018, doi: 10.1109/ACCESS.2018.2887095

[19] D. Fauri, D. R. dos Santos, E. Costante, J. den Hartog, S. Etalle, and S. Tonetta, "From system specification to anomaly detection (and back)," in *Proc. CPS*, Dallas, TX, 2017, doi: 10.1145/3140241.3140250

[20] J. A.Yaacoub, O. Salman, H. N. Noura, N. Kaaniche, A. Chehab, M. Malli, "Cyber-physical systems security: Limitations, issues and future trends," *Microprocessors and Microsystems*, vol. 77, 103201, Sept. 2020, doi: 10.1016/j.micpro.2020.103201

[21] H. Alemzadeh, D. Chen, X. Li, T. Kesavadas, Z. T. Kalbarczyk, and R. K. Iyer, "Targeted attacks on teleoperated surgical robots: Dynamic model-based detection and mitigation," in *Proc. DSN*, Toulouse, France, 2016, doi: 10.1109/DSN.2016.43

[22] T. Zebin, S. Rezvy, and Y. Luo, "An Explainable AI-based intrusion detection system for DNS over HTTPS (DoH) attacks," *IEEE Trans. Inf. Forensics Security*, vol. 17, pp. 2339–2349, Jun. 2022, doi: 10.1109/TIFS.2022.3183390

[23] C. Hwang and T. Lee, "Explainable sensor fault detection in the ICS anomaly detection system," *IEEE Access*, vol. 9, pp. 140470–140486, Oct. 2021, doi: 10.1109/ACCESS.2021.3119573

[24] H. Moraliyage, V. Sumanasena, D. De Silva, R. Nawaratne, L. Sun, and D. Alahakoon, "Multimodal classification of onion services for proactive cyber threat intelligence using explainable deep learning," *IEEE Access*, vol. 10, pp. 56044–56056, May 2022, doi: 10.1109/ACCESS.2022.3176965

[25] H. Suryotrisongko, Y. Musashi, A. Tsuneda, and K. Sugitani, "Robust botnet DGA detection: Blending XAI and OSINT for cyber threat intelligence sharing," *IEEE Access*, vol. 10, pp. 34613–34624, Mar. 2022, doi: 10.1109/ACCESS.2022.3162588

[26] H. Li, J. Wu, H. Xu, G. Li, and M. Guizani, "Explainable intelligence-driven defense mechanism against advanced persistent threats: A joint edge game and AI approach," *IEEE Trans. Dependable and Secure Comput.*, vol. 19, no. 2, pp. 757–775, Nov. 2021, doi: 10.1109/TDSC.2021.3130944

[27] S. Mane and D. Rao, "Explaining network intrusion detection system using explainable AI framework," arXIV, Mar. 2021, doi: 10.48550/arXiv.2103.07110

[28]R. Hamon, H. Junklewitz, I. Sanchez, G. Malgieri, and P. De Hert, "Bridging the gap between AI and explainability in the GDPR: Towards trustworthiness-by-design in automated decision-making," *IEEE Comput. Intell. Mag.*, vol. 17, no. 1, pp. 72–85, Feb. 2022, doi: 10.1109/MCI.2021.3129960
[29]R. V. Patil, N. P. Ambritta, P. N. Mahalle and N. Dey, "Medical cyber-physical systems in society 5.0: Are we ready?" *IEEE Trans. Technol. Society*, vol. 3, no. 3, pp. 189–198, Jun. 2022, doi: 10.1109/TTS.2022.3185396
[30]D. Gunning, E. Vorm, J. Y. Wang, and M. Turek, "DARPA's explainable AI (XAI) program: A retrospective," *Applied AI Letters*, vol. 2: e61, doi: 10.1002/ail2.61

Nancy Ambritta P. received her B.S. in Computer Science and Engineering from Anna University, Tamil Nadu, India in 2010 and her M.S. in Computer Engineering from Smt. Kashibai Navale College of Engineering, Pune, India in 2015. Her research interests are cloud security, future Internet and explainable AI.

She has five years of experience in the software engineering field and three years of experience in teaching and research. She is currently working as a senior data engineer at Glareal Software Solutions Pvt. Ltd. and is a research scholar at Vishwakarma Institute of Information Technology, Pune.

Ms Ambritta has published research articles in international journals and conferences. She has co-authored a book published by Taylor and Francis group. She is a reviewer for the Springer Journal of Wireless Personal Communications.

Parikshit N. Mahalle (Senior Member, IEEE) received his Ph.D from Aalborg University, Denmark and continued as Post Doc Researcher at CMI, Copenhagen, Denmark.

He is currently a Professor and Head of Department of Artificial Intelligence and Data Science at Vishwakarma Institute of Information Technology, Pune, India. He has 23 years of teaching and research experience.

Prof Mahalle has 9 patents, 200 research publications and authored/edited 43 books with Springer, CRC Press, Cambridge University Press, etc. He is editor in chief for IGI Global – International Journal of Rough Sets and Data Analysis, Inter-science International Journal of Grid and Utility Computing, member-Editorial Review Board for IGI Global – International Journal of Ambient Computing and Intelligence. His research interests are Machine Learning, Data Science, Algorithms, Internet of Things, Identity Management and Security.

Rajkumar V. Patil received his B.S. in Engineering (Computer Engineering) in 2018 from Savitribai Phule Pune University, Pune, India, and his M.S. in Computer Engineering in 2020 from Smt. Kashibai Navale College of Engineering, Pune, India.

He was appointed as an assistant professor at NBN Sinhgad School of Engineering, Pune, and is presently an assistant professor at Vishwakarma Institute of Information Technology, Pune, as well as a research scholar at the same institution.

Mr Patil's areas of interest include cyber physical systems for healthcare, trust management, machine learning, web technologies, and explainable AI. He has published research articles and book chapters in international journals. He reviews for several international journals.

Nilanjan Dey (Senior Member, IEEE) received his B.Tech. in Information Technology from the West Bengal University of Technology in 2005, M.Tech. in Information Technology in 2011 from the same University, and Ph.D. in digital image processing in 2015 from Jadavpur University, India.

In 2011, he was appointed as an Assistant Professor in the Department of Information Technology at JIS College of Engineering, Kalyani, India, followed by Bengal College of Engineering College, Durgapur, India, in 2014. He is now employed as an Associate Professor in the Department of Computer Science and Engineering Techno International New Town, Kolkata, India. His research topic is signal processing, machine learning, and information security.

Dr. Dey is an Associate Editor of IET Image Processing and is currently the Editor-in-Chief of the International Journal of Ambient Computing and Intelligence, and Series co-editor of Springer Tracts of Nature-Inspired Computing.

Rubén González Crespo (Senior Member, IEEE) received his PhD in Computer Science Engineering from the Pontifical University of Salamanca, Spain in 2008.

Currently, he is Vice-Chancellor of Academic Affairs and Faculty from UNIR and Global Director of Engineering Schools from PROEDUCA Group. He is an advisory board member for the Ministry of Education at Colombia and evaluator from the National Agency for Quality Evaluation and Accreditation of Spain (ANECA). He is a member of different committees at ISO Organization.

Prof. Crespo has published more than 200 papers in indexed journals and congresses.

R. Simon Sherratt (Fellow, IEEE) received the B.Eng. from Sheffield City Polytechnic in 1992, both the M.Sc. in 1993, and Ph.D. in 1996 from The University of Salford.

In 1996, he was appointed as a Lecturer in Electronic Engineering with the University of Reading, where he is currently a Professor of Biosensors. His research area is wearable devices, mainly for healthcare and emotion detection.

Eur Ing Professor Sherratt was awarded the 1st place IEEE Chester Sall Award in 2004, 2nd place in 2014, 3rd place in 2015 and 3rd place in 2016 for best papers in the IEEE TRANSACTIONS ON CONSUMER ELECTRONICS. He is currently Chair of the IEEE Masaru Ibuka Consumer Technology Award.