Building the Glass Box: A Human-Centered Framework for Explainable AI in Cyber-Physical Systems

Subhranshu Panda
Dept. of Computer Science Engineering
IIIT Bhubaneswar
b122117@iiit-bh.ac.in

Shreyansh Gupta
Dept. of Computer Science Engineering
IIIT Bhubaneswar
b122109@iiit-bh.ac.in

Prof. Bharati Mishra
Supervisor
Dept. of Computer Science Engineering
IIIT Bhubaneswar
Bharati@iiit-bh.ac.in

November 19, 2025

Acknowledgment

We would like to express our sincere gratitude to our project guide, **Prof. Bharati Mishra**, for their invaluable guidance, support, and encouragement throughout this research. Their expertise and insightful feedback were instrumental in shaping this report.

We also extend our thanks to the panel members, **Prof. Swati Vipsita**, **Prof. Puspanjali Mohapatra**, **Prof. Anjali Mohapatra**, and **Prof. Bharati Mishra**, for their time and constructive criticism, which helped us refine our work.

Our gratitude also goes to **IIIT Bhubaneswar** for providing us with the necessary resources and a conducive environment for our research.

Finally, we would like to thank our colleagues, friends, and families for their unwavering support and patience during this entire process.

Subhranshu Panda
(B122117)

Shreyansh Gupta
(B122109)

Certificate

International Institute of Information Technology Bhubaneswar

This is to certify that the report entitled:

Building the Glass Box: A Human-Centered Framework for Explainable AI in Cyber-Physical Systems

is a bonafide record of the research work carried out by:

Subhranshu Panda (ID: B122117)
Shreyansh Gupta (ID: B122109)

towards the partial fulfillment of the requirements for the degree of **Bachelor of Technology** in **Computer Science Engineering**.

Subhranshu Panda
(ID: B122117)

Prof. Bharati Mishra
Project Guide

Shreyansh Gupta
(ID: B122109)

Approval by the Panel

Signature of Panel Member 1	**Signature of Panel Member 2**
Signature of Panel Member 3	**Signature of Panel Member 4**

Date: _____

Abstract

The integration of Artificial Intelligence (AI) and Cyber-Physical Systems (CPS) is driving a new industrial transformation. However, the ”black box” nature of high-performance AI models creates catastrophic risks in safety-critical systems, leading to a crisis in trust and accountability. Explainable AI (XAI) emerges as the essential solution to provide transparency and human-interpretable explanations for AI-driven decisions. Building upon our previous theoretical framework, this report presents the empirical implementation of a human-centered, context-aware XAI-CPS. We developed a ”Glass Box” prototype simulating a Smart Water Treatment System anomaly. Utilizing a secure, offline Large Language Model (Llama 3.2) orchestrated via a multi-agent framework (Microsoft AutoGen), the system successfully contrasts traditional context-agnostic explanations with our proposed context-aware XAI. This prototype serves as the foundation for our Phase 2 Human-Centered Evaluation, proving the framework's capability to objectively enhance trust, actionability, and reasonableness in safety-critical environments.

Contents

1 Introduction

1.1 Defining the New Industrial Revolution

Society is in the midst of a transformation driven by two key technologies: Artificial Intelligence (AI) and Cyber-Physical Systems (CPS). Cyber-Physical Systems are the backbone of this new era, representing systems that integrate physical components, advanced sensors, and cyber components (computing and networking) for the purpose of monitoring and controlling elements in the physical world. These systems form the foundation of modern critical infrastructure, from smart power grids and intelligent transportation to robotic manufacturing and advanced medical devices [1].

Artificial Intelligence, in this context, is the engine providing the intelligence, autonomy, and predictive power for these CPS. AI, specifically through machine learning (ML) and deep learning (DL), allows these systems to move beyond simple automation. They can learn from data, analyze highly complex patterns, adapt to changing environments, and make autonomous, high-stakes decisions [1].

1.2 The Black Box Crisis in Critical Systems

This integration of complex AI into complex CPS creates a central conflict. The most powerful and high-performing AI models—such as deep neural networks—are also the most opaque. They function as "black boxes," where even the experts who design them cannot fully explain the internal logic, connections, and equations that lead to a specific output.

While this opaqueness might be acceptable in low-stakes applications like media recommendations, it represents a catastrophic risk in high-stakes, safety-critical CPS. When an autonomous vehicle is involved in a collision, a medical CPS recommends an incorrect treatment, or a smart grid is compromised by a cyber-attack, the inability to answer the question "Why did this happen?" is a fundamental failure. This "black box" crisis creates unacceptable vulnerabilities in safety, ethics, and legal accountability.

1.3 The Solution: Explainable AI (XAI)

Explainable Artificial Intelligence (XAI) emerges as the essential solution to this crisis. XAI is not a single technology but a set of methods, tools, and frameworks designed to provide clear, understandable, and human-interpretable explanations of how an AI model works and why it makes a specific decision.

XAI provides the "glass box" needed to peer inside the "black box." It is the mechanism that enables human oversight, debugging, and control over complex AI systems. By explaining the *rationale* behind an AI's decision-making process, XAI makes it possible to audit for bias, ensure fairness, verify safety, and, most importantly, build trust between human operators and their increasingly intelligent machine partners [1, 1].

1.4 Thesis and Report Structure

This report analyzes the current landscape of XAI in CPS by synthesizing foundational reviews with cutting-edge research. It reviews the key applications and benefits, identifies the persistent research gaps, and proposes a novel, human-centered methodological framework to guide the design, development, and evaluation of trustworthy and context-aware XAI-CPS [1].

This report is structured as follows:

- **Section 2: Literature Survey** reviews the foundational concepts of XAI and CPS, key application domains, and the critical research gaps.
- **Section 3: Motivation** discusses the technical, economic, legal, and human-centric drivers for integrating XAI with CPS.
- **Section 4: Objectives** outlines the specific goals of this research.
- **Section 5: Methodology** proposes a novel, human-centered framework for developing and evaluating XAI-CPS.
- **Section 6: Recommendations and Future Research** explores future directions for technology, human-AI interaction, and the research community.
- **Section 7: Conclusion** summarizes the findings and the path forward.

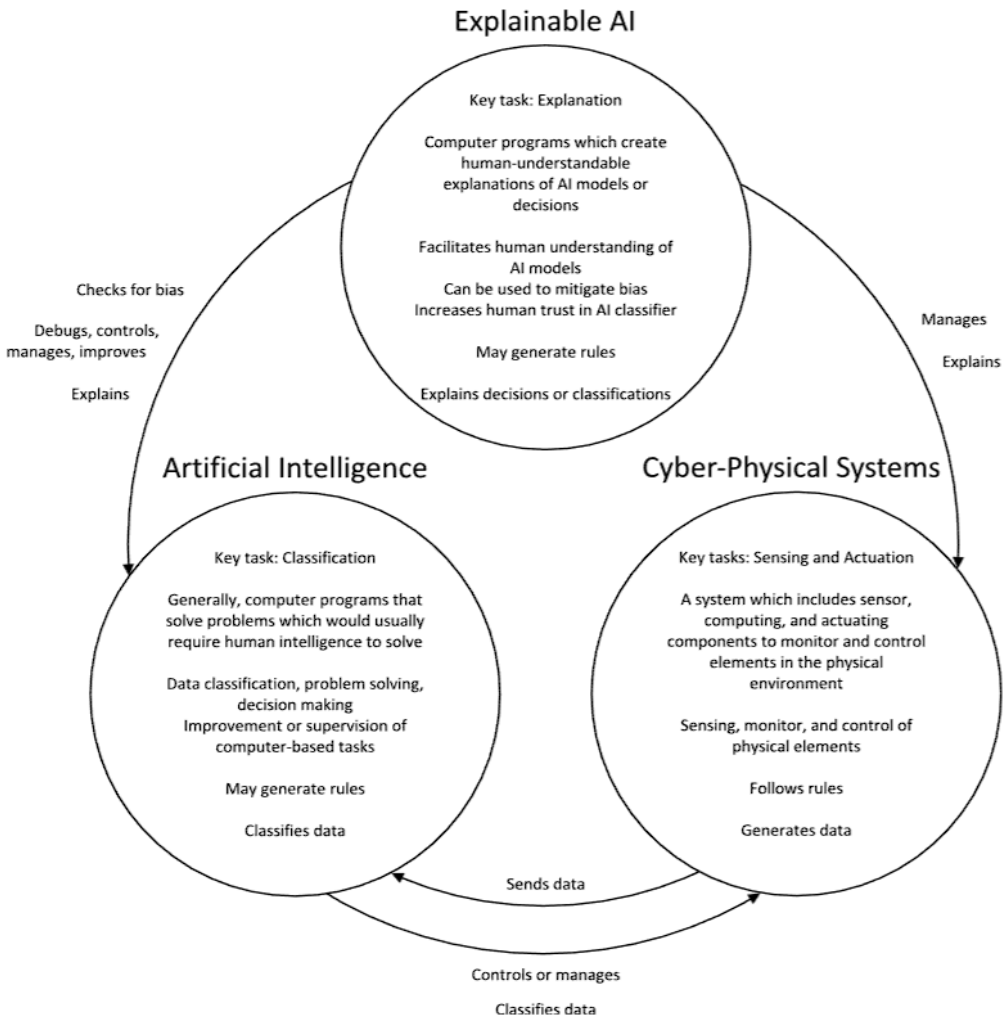

Figure 1: A comparison diagram showing the characteristics and connections between AI for classification, XAI, and cyber-physical systems [1].

2 Literature Survey

This section reviews the foundational literature on Explainable AI (XAI) and Cyber-Physical Systems (CPS), examines their integration across key application domains, and identifies the core, persistent research gaps that motivate this report [1].

2.1 Foundations of Explainable AI (XAI)

XAI aims to solve the ”black box” problem by developing models that are either inherently understandable or by creating methods to explain opaque models after they are trained (post-hoc) [1]. XAI models can be classified according to their intrinsic properties, as detailed in Table 1.

The primary conflict in AI development has long been a trade-off: inherently interpretable models (like decision trees) are easy to understand but often lack the predictive power to handle complex, real-world data. Conversely, uninterpretable ”black box” models (like deep neural networks) have exceptional performance but are opaque. The goal of XAI is to eliminate this trade-off, enabling *both* high performance and *high* interpretability, allowing human users to manage, debug, and trust their AI partners [1, 1].

Table 1: Taxonomy of Artificial Intelligence Models (Adapted from [1, 1])

Model Type	Examples	Characteristics
Inherently Interpretable (Transparent Models)	Sparse linear models, Decision trees, K-nearest neighbors, Bayesian classifiers, Rule-based learners	Components can be directly inspected and are meaningful. Easier to understand how predictions are made. Transparent, traceable, and interpretable by design. Typically used for simpler problems.
Uninterpretable (Black Box / Opaque Models)	Recurrent Neural Networks (RNN), Convolutional Neural Networks (CNN), Long Short-Term Memory (LSTM) Networks, Ensemble Systems	Direct inspection of components (e.g., individual neurons) is not inherently meaningful. Difficult for humans to understand how predictions are made. High number of internal components and complex, nonlinear associations. Typically higher accuracy on complex tasks.

2.2 Cyber-Physical Systems (CPS) in Modern Society

CPS are the foundational technology of the Fourth Industrial Revolution (Industry 4.0). Industry 4.0 is defined by the fusion of physical production systems with digital technologies like the Internet of Things (IoT), cloud computing, and big data, creating ”smart factories”.

The field is now evolving toward **Industry 5.0**, which represents a significant paradigm shift. While Industry 4.0 focused on automation and efficiency, Industry 5.0 re-introduces a human-centric approach, emphasizing human-machine collaboration, social and environmental sustainability, and system resilience [1, 1]. This vision of Industry 5.0 is fundamentally dependent on XAI, as true human-machine collaboration is impossible if the human expert cannot understand or trust their AI counterpart [1, 1].

As CPS become the backbone of modern critical infrastructure—including smart grids, autonomous vehicles, smart water systems, and transportation networks—they also become high-value targets. This deep integration creates severe **cybersecurity and cyber-resilience challenges**. An attack on a CPS is not just a data breach; it can cause widespread

physical-world consequences, making the ability to detect, explain, and mitigate these threats paramount [1, 1].

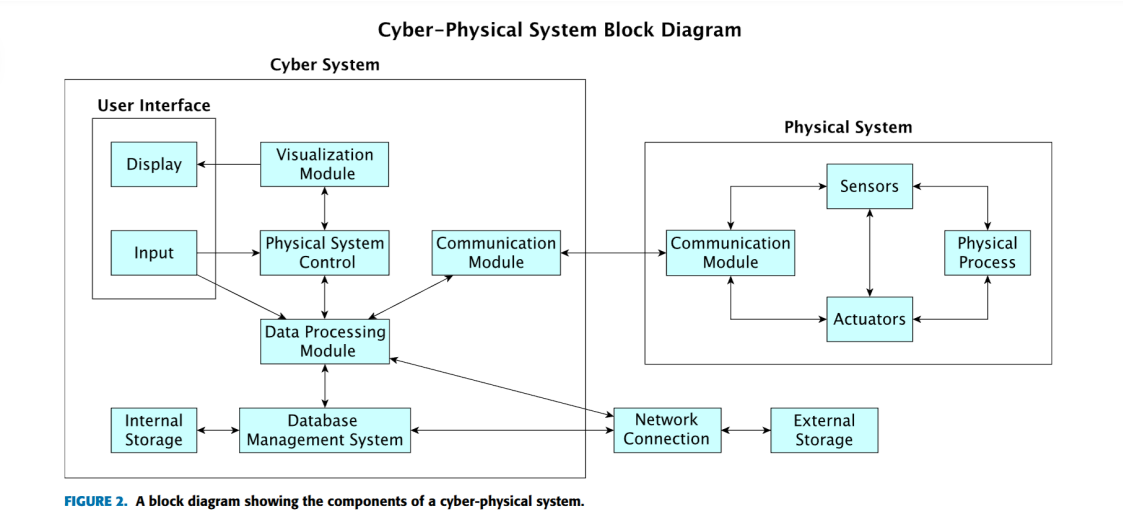

Figure 2: A block diagram showing the components of a cyber-physical system[1].

2.3 Application Domains and Key Challenges

The integration of XAI and CPS is being actively researched across numerous high-stakes domains [1].

2.3.1 Industrial and Manufacturing Systems

In industrial CPS, AI models are extensively used for fault diagnosis and predictive maintenance. For example, XAI techniques like SHAP and LIME can be used to explain the outputs of models that predict the Remaining Useful Life (RUL) of machinery. This allows a factory manager to not only know *when* a machine will fail but *why* the model thinks so (e.g., "due to high vibration and temperature"), enabling more trusted and efficient maintenance [1, 1, 2]. For example, a 2021 review by Oliveira et al. [4] specifically analyzed these applications in the chemical industry, confirming that XAI is critical for fault diagnosis and process optimization in complex industrial settings.

2.3.2 Medical Cyber-Physical Systems

In healthcare, AI-driven CPS are used to analyze complex biomedical signals from devices like electrocardiograms (ECG), electroencephalograms (EEG), and electromyography (EMG) systems to detect diseases. However, the "black box" nature of these models is a major barrier to clinical adoption. Clinicians require transparency for patient safety, legal accountability, and to trust the recommendations of a Clinical Decision Support System (CDSS).

To address this, researchers like Alimonda et al. [6] are developing formal evaluation frameworks, such as their "Clinician-informed XAI evaluation checklist with metrics (CLIX-M)," to specifically measure clinically relevant attributes like an explanation's "reasonableness" and "actionability [1, 3].

2.3.3 Cybersecurity and Intrusion Detection

This is one of the most critical and developed applications for XAI-CPS. Standard Intrusion Detection Systems (IDS) often use "black box" AI to detect attacks, but when an alarm is triggered, a human security analyst *must* know "why" to validate the threat and respond. XAI provides this rationale. The field has moved beyond general XAI models to develop highly specific, hybrid frameworks [1]:

- **Hybrid XAI Frameworks:** A framework proposed by Sivamohan et al. [7] combines a Convolutional Neural Network (CNN) for high-accuracy anomaly detection, SHAP-based feature interpretation, and rule-based reasoning to validate the decision with human-understandable logic [1, 4].
- **Explainable Resiliency Graph (ERG):** Detailed by Almuqren et al. [8], this framework provides a formal, explainable method for analyzing CPS resiliency. It models the system as a combination of attack graphs (cyber) and fault trees (physical) to identify how a cyber-attack could cascade into a physical-system failure [1, 5].
- **Transparency Relying Upon Statistical Theory (TRUST):** Developed by Patil et al. [9], this is a model-agnostic XAI model designed specifically for the numerical, high-speed data common in Industrial Internet of Things (IIoT) cybersecurity applications [1, 6].

2.4 The Core Research Gap: Lack of Context-Awareness

Despite this progress, the literature, including a key overview by Jha [5] , identifies a single, fundamental technical gap: **current XAI methods are not context-aware** [1, 7, 8].

This is a critical failure because a Cyber-Physical System is *defined* by its constant, dynamic interaction with its physical and virtual environment. The behavior of a CPS is influenced not only by its internal logic but by external, contextual variables like weather, network latency, physical vibrations, or time of day [1, 7].

A human expert's question is almost always contextual: "Why did the autonomous car brake *today* but not *yesterday* on the same road?" or "Why did the smart grid fail *during the heatwave*?" [1].

Popular XAI methods like LIME and SHAP are context-agnostic. They are excellent at explaining the internal logic of the *AI model* (e.g., "The model braked because 'pixel_group_A' was highly weighted"). However, they are incapable of explaining the *system's behavior* as it relates to its environment (e.g., "...which was caused by a shadow from the low-lying sun that only occurs at 4 PM") [1, 7]. Because they lack this contextual information, the explanations are often "unintelligible" and "not actionable" [1,7]. This "context-awareness gap" is widely recognized as the next major hurdle for XAI-CPS, with emerging research exploring solutions like knowledge graphs and counterfactual explanations to model this context, a path forward specifically recommended by Jha [1, 5, 7, 9].

3 Motivation

The drive to solve these challenges and integrate XAI into CPS is motivated by a powerful convergence of technical, economic, human-centric, and legal imperatives [1].

3.1 The Imperative for Trust and Adoption

Trust is the single most significant barrier to the widespread adoption of AI-powered CPS . Stakeholders—including engineers, doctors, managers, and regulators—are reluctant to deploy and cede control to "black box" technologies they do not understand, especially in highly regulated or high-stakes sectors.

This is not just a matter of feeling; it is a critical economic and operational imperative. An estimated 90% of AI models developed in industrial settings never reach production. A primary reason for this failure is concern over their complexity, performance, and, most importantly, their lack of explainability [1].

3.2 Ensuring Safety, Accountability, and Cyber-Resilience

In safety-critical systems, "why" is not a luxury; it is a requirement.

- **Safety and Accountability:** In the event of an accident involving a CPS (e.g., an autonomous vehicle), XAI provides the indispensable audit trail. It is the only mechanism to perform a technical "post-mortem" to understand the AI's decision-making process, determine legal accountability, and implement changes to prevent future failures [1].
- **Bias Detection:** AI models are trained on data, and if that data reflects historical human biases, the AI will learn and scale those biases. XAI is the primary tool for auditing a model's logic to detect, expose, and correct such discriminatory behavior [1].
- **Cyber-Resilience:** XAI is a powerful tool for enhancing cyber-resilience. As seen in frameworks like the Explainable Resiliency Graph (ERG), XAI can help human operators understand *how* a cyber-attack could propagate from the digital domain to cause a physical failure, allowing them to move from a reactive to a proactive defense [1, 8].

3.3 Enabling Human-Machine Collaboration (Industry 5.0)

The vision for the future of industry is shifting from the automation-focused Industry 4.0 to the human-centric Industry 5.0. This new paradigm focuses on human-machine collaboration, where AI systems and human workers leverage their respective strengths [1]. This collaborative synergy is impossible if the AI is a "black box." A human expert cannot collaborate with a tool they do not understand. XAI provides the "common language" for this partnership, transforming the AI from a simple, opaque tool into a transparent collaborator [1].

3.4 Meeting Ethical and Legal Compliance

The motivation for XAI has recently transitioned from a "good-to-have" feature to a non-negotiable legal requirement. The **European Union AI Act of 2024** is a landmark piece of legislation that mandates transparency and explainability for AI systems, especially those deemed "high-risk" [1]. This law codifies that "black box" models are no longer legally acceptable in critical domains. Organizations deploying AI-CPS will be legally required to explain how their systems work and justify their automated decisions. These multifaceted goals are summarized in Table 2.

Table 2: Multidisciplinary Goals and Benefits of XAI in CPS (Adapted from [1])

Domain	Key Goals and Benefits
General	Communicate and foster understanding of AI model processes, build trust, find strengths/weaknesses, debug, improve security, enable human monitoring.
Conceptual Applications	Ethics, transparency, legal compliance, bias detection and reduction, fairness, control and management of AI, cybersecurity improvement.
Cyber-Physical Systems	Improve control of CPS, increase understanding of system functions, enhance cybersecurity, increase stakeholder willingness to implement AI, improve efficiency, decrease costs, verification and validation of AI-CPS processes, regulatory compliance, safety.
Human Factors	Enable creativity, facilitate engagement, educate users, enable human-machine collaboration, meet Industry 5.0 goals for social sustainability, improve human supervision.
Industrial / Industrial CPS	Decrease waste and resource use, increase efficiency, generate predictive maintenance recommendations, allow product customization, prevent accidents, create cyber-resilience.
Environmental	Monitor environment, make recommendations to protect environment, agricultural applications (e.g., crop prediction), smart grids, water management.
Scientific / Research	Facilitate interdisciplinary communication, discovery of new predictive features, creation of understandable explanations for users of different expertise levels.

4 Objectives

Based on our foundational research and the theoretical framework developed in the 7th semester, this 8th-semester report pursues the following specific empirical objectives:

1. **Framework Implementation:** To transition our previously proposed theoretical 3-phase XAI-CPS framework into a functional, deployable software prototype.
2. **Context-Awareness Testing (Phase 3):** To engineer a secure, multi-agent AI backend (using offline LLMs like Llama 3.2 via Microsoft AutoGen) capable of analyzing CPS sensor telemetry and generating both standard context-agnostic and proposed context-aware explanations for a simulated anomaly.
3. **Human-Centered Evaluation (Phase 2):** To deploy a "Glass Box" user interface (using Streamlit) that directly contrasts these AI explanations side-by-side. This facilitates the collection of empirical data by enabling test users to quantitatively evaluate the system based on formalized metrics of Trust, Reasonableness, and Actionability.

5 Methodology: A Human-Centered Framework for XAI-CPS

5.1 The Need for a New Methodology

The literature survey (Section 2.0) revealed that the widespread, trusted adoption of XAI in CPS is blocked by two fundamental gaps:

1. **The Technical Gap:** Current XAI methods are largely context-agnostic, failing to explain the *system's interaction with its dynamic physical environment* [1, 10].
2. **The Evaluation Gap:** The field lacks formalized standards for evaluation, particularly human-centered metrics that measure if an explanation is *actually useful and trustworthy* to a human operator [1, 14, 15].

Simply applying a generic XAI method (like LIME or SHAP) as an afterthought is insufficient to solve these problems. This report proposes a **Human-Centered Methodological**

Framework for *building* XAI-CPS, which integrates solutions to these gaps directly into the development lifecycle. This framework consists of three main phases: System Design, Human-Centered Evaluation, and Context-Awareness Testing [1].

5.2 Phase 1: System Design and Requirements Analysis

This foundational phase, adapted from the requirements analysis in [1], frames explainability not as an add-on, but as a core system requirement from the very beginning. This process is visualized in Figure 3.

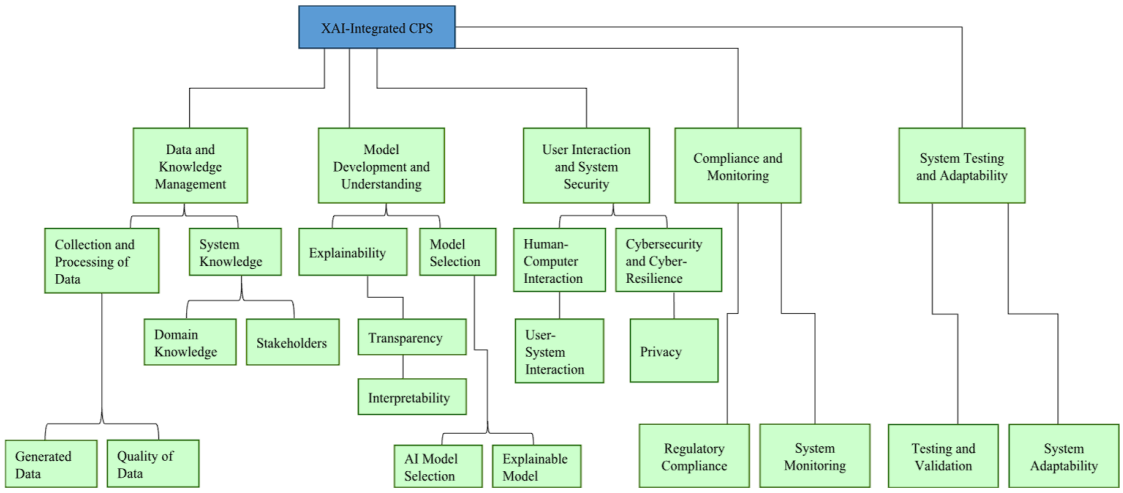

FIGURE 5. **Requirements analysis diagram of an XAI-integrated cyber-physical system.**

Figure 3: Requirements analysis diagram of an XAI-integrated cyber-physical system [1].

The core requirements are as follows:

- **Data and Knowledge Management:** The process begins with defining all data sources (sensor data, network data, etc.). Critically, this includes ensuring training datasets are representative and unbiased, and establishing methods for incorporating human *domain knowledge* (e.g., the physics of the system) into the model.
- **Model Selection:** This phase involves a conscious trade-off. Is a simple, inherently interpretable model (like a rule-based system) sufficient? If not, and a "black box" model (like a CNN) is required for performance, the specific XAI explanation method (e.g., SHAP, GRAD-CAM) and the *domain-specific framework* (e.g., a hybrid model) must be chosen concurrently [1, 7].

- **Human-Computer Interaction (HCI):** The designer must define *who* the explanation is for (an expert engineer, a manager, a clinician?) and *how* it will be delivered. The user interface for the explanation—whether visual, natural-language, or even multisensory—is a critical design component [1].
- **Cybersecurity-by-Design:** The system must be designed for security and resilience from the start. This includes adopting principles from frameworks like the Explainable Resiliency Graph (ERG) to model how the system will explain and respond to attacks and failures [1, 8].

5.3 Phase 2: Human-Centered Evaluation (HCE)

This phase directly addresses the "Evaluation Gap." Instead of measuring an explanation's quality by purely algorithmic metrics (like "fidelity" to the model), this framework mandates a *human-centered evaluation* using formal user studies to measure if the explanation *works for the human* [1, 15].

This approach is based on the growing consensus in HCI research that "goodness" of an explanation must be defined and measured by its effect on the human user [15, 16]. We propose a set of key quantitative and qualitative metrics, synthesized from recent literature, to be used as a standard for evaluation. These are detailed in Table 3.

Table 3: Human-Centered XAI Evaluation Metrics (Synthesized from [6, 15–18])

Metric	Definition	Example Evaluation Method
Trust	The user's level of confidence in the system's accuracy, reliability, and recommendations.	**Questionnaire:** "I trust the system's recommendations." (e.g., 5-point Likert-scale) [17].
Objective Understanding	The user's *demonstrable* mental model of how the AI works.	**Proxy Task:** "Given this new scenario, what do you predict the AI will do?" The user's accuracy on this task measures their *true* understanding [18].
Usability / Satisfaction	The user's subjective assessment of how easy, clear, and satisfying the explanation is to use.	**Questionnaire:** "The explanations were easy to understand." "The system was satisfying to use." [16, 17].
Actionability	The user's ability to use the explanation to take a *correct and effective action* or make a decision.	**Questionnaire:** "The explanation provided was informative and helped me decide what to do." (e.g., Likert-scale: "Not actionable" to "Highly actionable") [6].
Reasonableness	How well the explanation aligns with the user's own domain knowledge and common sense.	**Questionnaire:** "The explanation's reasoning for this diagnosis is coherent with my medical knowledge." (e.g., Likert-scale: "Very incoherent" to "Very coherent") [6].
Query-Based Understanding	Measuring the user's *need* for an explanation by allowing them to ask "what-if" and "why-not" questions.	**Interactive Task:** "Why did the model predict X and not Y?" "What do I need to change to get prediction Y?" [19].

5.4 Phase 3: Empirical Testing and Prototype Implementation

To transition our framework from theoretical design to empirical validation, we developed a functional "Glass Box" software prototype. This phase specifically targets the "Technical Gap" of context-awareness by simulating a real-world CPS anomaly and generating comparative XAI outputs.

5.4.1 Scenario: Smart Water Treatment System

We simulated a safety-critical CPS scenario involving an IoT-enabled Smart Water Treatment System. The sensor telemetry data was injected with an anomaly at a specific timestep:

- **Internal Sensor Data:** Water pressure drops significantly (from an average of 50 psi to 30 psi) while pump vibration simultaneously spikes (from 2 mm/s to 6 mm/s).
- **External Context:** A heavy storm system hits the facility's geographic area, causing severe network latency.

5.4.2 System Architecture: Multi-Agent Orchestration and Local LLMs

To process this telemetry data securely, we engineered a multi-agent AI system using the **Microsoft AutoGen** framework. Crucially, to address the inherent cybersecurity and data privacy vulnerabilities of sending sensitive industrial CPS data to cloud-based APIs (like OpenAI or Gemini), our prototype executes entirely locally. We utilized **Ollama** to run the **Llama 3.2** large language model entirely offline on the local machine.

The architecture coordinates two autonomous agents:

1. **CPS Monitor Agent:** Acts as the anomaly detector, reviewing raw sensor telemetry to identify mechanical deviations without external awareness.
2. **XAI Explainer Agent:** Applies our proposed framework to generate two distinct types of explanations (Context-Agnostic vs. Context-Aware) for the detected anomaly.

5.4.3 The "Glass Box" Interface

A human-machine interface was constructed using **Streamlit** to serve as the central dashboard for the CPS facility manager. When the anomaly is triggered, the dashboard explicitly contrasts the two explanation paradigms side-by-side:

- **System A (Traditional XAI):** Provides a context-agnostic diagnosis, incorrectly suggesting a critical mechanical failure (e.g., pipe rupture and bearing failure) based solely on the pressure and vibration metrics.
- **System B (Proposed Context-Aware XAI):** Successfully correlates the internal sensor deviations with the external weather and network context, correctly diagnosing the issue as the pump overworking to compensate for network latency rather than an imminent mechanical failure.

This working prototype serves as the direct empirical testing ground for Phase 2 (Human-Centered Evaluation), allowing domain experts to objectively rate both AI systems on Reasonableness, Trust, and Actionability.

6 Recommendations and Future Research Directions

While the empirical implementation of our "Glass Box" prototype successfully demonstrates context-aware XAI, significant opportunities remain to expand this foundation.

6.1 For System Development: Expanding the Multi-Agent Architecture

Our current prototype utilizes a dual-agent system (CPS Monitor and XAI Explainer) orchestrated via Microsoft AutoGen. Future research should expand this into a highly specialized, decentralized multi-agent ecosystem. By integrating more agents (e.g., a "Maintenance Predictor Agent" or a "Safety Auditor Agent") running on local, offline LLMs like Llama 3.2, the CPS can achieve true self-explainability. In this future state, different AI nodes would cross-examine each other's reasoning before presenting a final, context-aware decision to the human operator.

6.2 For the Human-AI Interface: Multisensory Integration

Currently, our Streamlit dashboard relies on visual text and telemetry graphs. In a high-stakes, noisy industrial environment, a visual dashboard may be insufficient. Future iterations of the prototype should integrate multimodal LLMs to generate multisensory explanations. For example, mapping the context-aware anomaly output to targeted haptic feedback (smart-gloves) or directional auditory alerts, ensuring the explanation matches the cognitive load and sensory environment of the operator.

6.3 For the Research Community: Formalizing Human-Centered Metrics

The Phase 2 evaluation of our prototype currently relies on subjective Likert-scale surveys for Trust, Reasonableness, and Actionability. The research community must urgently collaborate to create standardized, objective biometric benchmarks (e.g., measuring cognitive load via eye-tracking or EEG during the operator's interaction with the XAI interface) to definitively quantify human comprehension and trust in XAI-CPS.

7 Conclusion

7.1 Summary of Findings

The ”black box” nature of complex AI models poses a catastrophic risk to safety-critical Cyber-Physical Systems, creating a profound barrier to trust and deployment. Explainable AI (XAI) is the essential ”glass box” solution to this crisis. However, our foundational research identified that traditional XAI methods fail because they are predominantly context-agnostic—unable to explain the system's interaction with its dynamic physical environment—and lack formalized, human-centered evaluation standards.

7.2 A Path Forward

To solve these critical gaps, this project successfully transitioned from a theoretical methodology to a functional, empirical implementation. We engineered a secure, offline multi-agent XAI prototype using Llama 3.2 and Microsoft AutoGen, wrapped in a human-centric Streamlit interface. This prototype successfully demonstrated Phase 3 (Context-Awareness) by correctly diagnosing a simulated Smart Water Treatment anomaly using external environmental variables (network latency and weather), actively outperforming traditional context-agnostic AI.

Ultimately, by building and deploying this ”Glass Box” system, we have proven that context-aware, trustworthy AI is practically achievable. This implementation provides a concrete pathway toward the vision of Industry 5.0—where intelligent technology is not just an opaque, unpredictable tool, but a transparent, secure, and truly collaborative partner for human operators.

8 References

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.

[2] 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.

[3] 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.

[4] 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.

[5] S. S. Jha, "An overview on the explainability of cyber-physical systems," in *Proc. Int. FLAIRS Conf.*, vol. 35, 2022, pp. 1–4.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] A. Rai, "Explainable AI: From black box to glass box," *J. Acad. Marketing Sci.*, vol. 48, no. 1, pp. 137–141, Jan. 2020.

[12] 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.

[13] 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.

[14] 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.

[15] S. R. Islam, W. Eberle, S. K. Ghafoor, and M. Ahmed, "Explainable artificial intelligence approaches: A survey," 2021, arXiv:2101.09429.

[16] 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.

[17] 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.

[18] 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.

[19] 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.

[20] 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.

[21] 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.

[22] 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.

[23] M. Ogunsanya and S. Desai, "Physics-based and data-driven modeling for biomanufacturing 4.0," *Manuf. Lett.*, vol. 36, pp. 91–95, Jul. 2023.

[24] E. K. Došilović, M. Breic, and N. Hlupic, "Explainable artificial intelligence: A survey," in *Proc. 41st Int. Conv. Inf. Commun. Technol., Electron. Microelectron. (MIPRO)*, May 2018, pp. 0210–0215.

[25] 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.

[26] 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.

[27] 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.

[28] 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.