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

Project Report Submitted
in partial fulfillment of the requirements for the award of the

Bachelor of Technology

in

Computer Science and Engineering

by

Subhranshu Panda
(B122117)

Shreyansh Gupta
(B122109)

Under the supervision of
Prof. Bharati Mishra

Department of Computer Science and Engineering
International Institute of Information Technology, Bhubaneswar
February 2026

Approval of the Viva-Voce Board

Date: ________________, 2026

Certified that the report entitled "Building the Glass Box : A Human-Centered Framework for Explainable AI in Cyber-Physical Systems" submitted by Subhranshu Panda (B122117) and Shreyansh Gupta (B122109) to International Institute of Information Technology Bhubaneswar in partial fulfillment of the requirements for the award of the Bachelor of Technology in Computer Science Engineering under the BTech Programme has been accepted by the examiners during the viva-voce examination held today.

(Supervisor)	(Panel Head)
(Internal Examiner 1)	(Internal Examiner 2)

Certificate

This is to certify that the report entitled "Building the Glass Box : A Human-Centered Framework for Explainable AI in Cyber-Physical Systems" submitted by Subhranshu Panda (B122117) and Shreyansh Gupta (B122109) to International Institute of Information Technology Bhubaneswar is a record of bonafide project work under my supervision, and the report is submitted for the award of the Bachelor of Technology in Computer Science and Engineering.

Prof. Bharati Mishra
Supervisor
Department of Computer Science and Engineering
IIIT Bhubaneswar

Declaration

We, Subhranshu Panda (B122117) and Shreyansh Gupta (B122109), hereby declare that this project report entitled "Building the Glass Box : A Human-Centered Framework for Explainable AI in Cyber-Physical Systems" submitted to International Institute of Information Technology Bhubaneswar is a record of an original work done by us under the guidance of Prof. Bharati Mishra, and this project work has not formed the basis for the award of any Degree or diploma/associate ship/fellowship and similar project if any.

Subhranshu Panda
(B122117)

Shreyansh Gupta
(B122109)

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 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)

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 The Rise of AI in Cyber-Physical Systems

Cyber-Physical Systems (CPS) represent the profound integration of computation, networking, and physical processes. They form the backbone of critical modern infrastructure, including smart manufacturing, autonomous transportation, and healthcare systems [2, 3]. The advent of Industry 4.0 and the ongoing transition towards Industry 5.0 heavily rely on deploying advanced Deep Learning models within these networks to achieve predictive maintenance, real-time optimization, and autonomous decision-making [26].

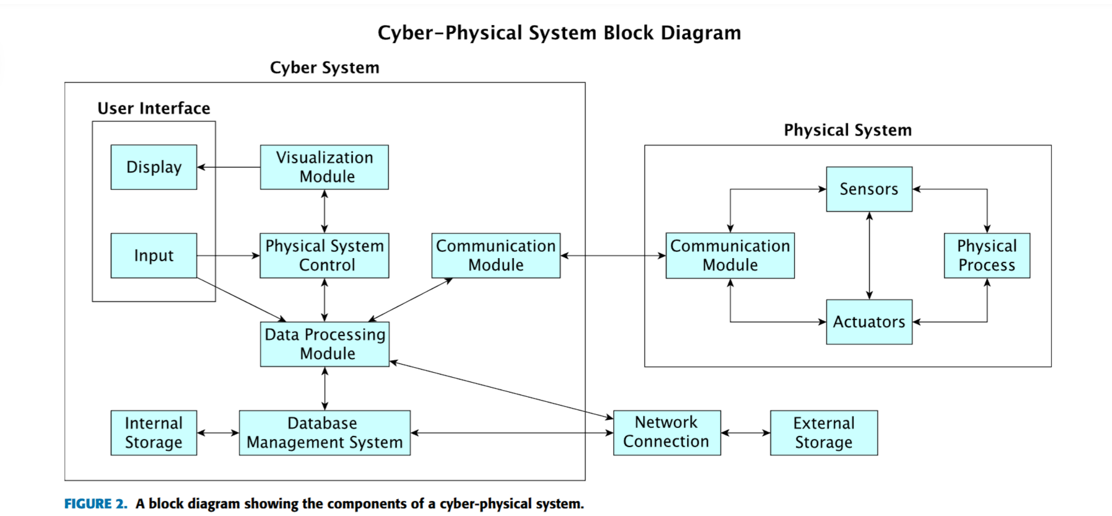

Figure 1: A block diagram showing the components of a cyber-physical system [1].

1.2 The "Black Box" Crisis

Despite their high predictive accuracy, state-of-the-art deep learning models inherently function as opaque "black boxes." While this opacity may be acceptable in low-risk applications, it introduces catastrophic risks in safety-critical CPS environments [25]. When an autonomous system makes a critical error or an industrial plant experiences an unpredicted anomaly, human operators cannot answer the fundamental question: *"Why did the AI make this decision?"* This lack of algorithmic transparency creates a severe deficit in trust, safety validation, and operational accountability.

1.3 The Need for Explainable AI (XAI)

Explainable AI (XAI) has emerged as the essential solution to this crisis. XAI provides the necessary "glass box" methodologies to produce human-interpretable explanations for complex AI behaviors, transforming opaque algorithms into accountable, auditable tools [24]. However, successfully implementing XAI in CPS requires moving beyond mere algorithmic transparency to true environmental context-awareness, which forms the core focus of this project's empirical implementation.

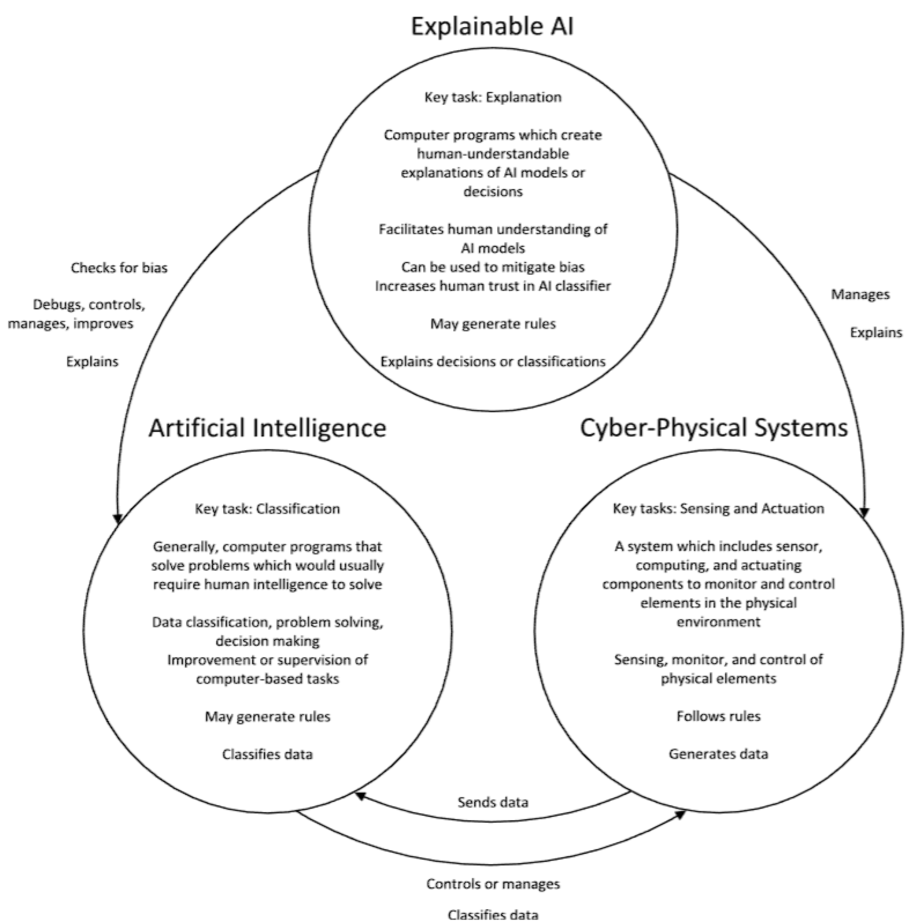

Figure 2: A comparison diagram showing the characteristics and connections between AI for classification, XAI, and cyber-physical systems [1].

2 Literature Survey

The integration of Explainable AI (XAI) within Cyber-Physical Systems (CPS) has garnered significant academic attention, primarily focusing on resolving the ”black box” nature of deep learning models in safety-critical domains such as smart manufacturing, healthcare, and autonomous grids. However, a critical review of the current state-of-the-art reveals a persistent over-reliance on traditional, context-agnostic XAI techniques (e.g., SHAP, LIME). These methods evaluate internal model features in a vacuum, completely ignoring the dynamic external physical environments in which these machines operate.

To identify the specific technical and methodological gaps addressed by our empirical prototype, Table 1 summarizes the key contributions and critical limitations of prominent recent literature in this domain.

Reference/Author	**CPS Domain**	**XAI Methodology**	**Critical Limitation / Gap**
Wickramasinghe et al. [1]	Industrial IoT	Unsupervised ML feature extraction	Lacks human-centric evaluation metrics; purely algorithmic focus.
Oliveira et al. [4]	Chemical Industry	Fault diagnosis and optimization	Context-agnostic; does not correlate mechanical faults with external environments.
Alimonda et al. [6]	Medical CPS	CLIX-M evaluation framework	Evaluated only in theoretical simulations; lacks empirical real-world testing.
Almuqren et al. [8]	CPS Security (IDS)	Feature weighting via SHAP/LIME	Explanations are isolated to network traffic, ignoring physical sensor states.
Our Proposed Work	Smart Water Treatment	**Multi-Agent Context-Aware LLMs**	**Bridges the gap by correlating internal telemetry with external environmental variables.**

Table 1: Summary of State-of-the-Art XAI in CPS and Identified Research Gaps

This comprehensive analysis confirms that while algorithmic transparency is improving, the industry desperately needs a *context-aware* framework evaluated through empirical, human-centered metrics—which directly motivates our 8th-semester prototype implementation.

3 Motivation

While the necessity of Explainable AI (XAI) in Cyber-Physical Systems (CPS) is well-documented [24, 25], a profound gap remains in how these explanations are practically generated. Current XAI frameworks are predominantly *context-agnostic* [1, 4]. They excel at highlighting which internal data features drove a model's prediction (e.g., mathematically flagging a high pump vibration), but they consistently fail to correlate these mechanical anomalies with the dynamic external physical environments (e.g., severe weather or network latency) that actually caused them. This isolation leads to false positives, unnecessary system downtime, and a rapid breakdown of human operator trust [26].

Motivated by our foundational theoretical research in the 7th semester, the primary driver for this 8th-semester implementation is to bridge the gap between algorithmic theory and operational reality. It is no longer sufficient to merely propose theoretical XAI models; the industry urgently requires functional, *context-aware* implementations that can be empirically tested. By developing a secure, multi-agent XAI prototype using offline Large Language Models, this project aims to prove that integrating external environmental variables into AI explanations directly and measurably enhances human trust, reasonableness, and actionability in industrial crisis management.

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

The widespread, trusted adoption of XAI in CPS is currently blocked by two fundamental gaps: a **Technical Gap** (lack of context-awareness) [1,10] and an **Evaluation Gap** (lack of formalized human-centered standards) [1,14,15]. To solve this, we propose and implement a three-phase methodological framework that integrates these solutions directly into the development lifecycle.

5.1 Phase 1: System Design and Requirements Analysis

This foundational phase frames explainability not as an add-on, but as a core system requirement from inception [1].

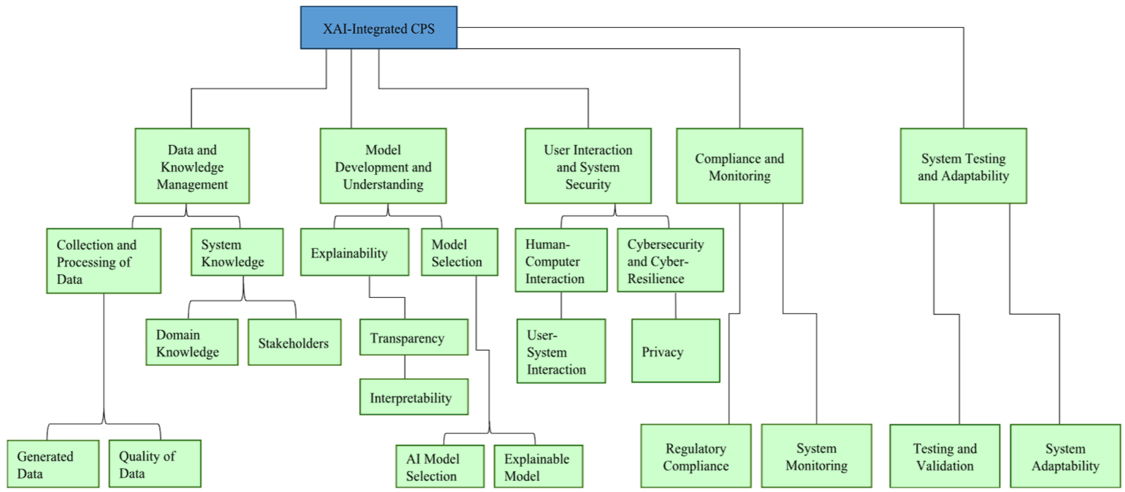

FIGURE 5. **Requirements analysis diagram of an XAI-integrated cyber-physical system.**

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

The design process mandates four critical pillars:

- **Data and Knowledge Management:** Ensuring representative datasets and establishing formal methods to incorporate human domain physics into the AI model.
- **Model Selection:** Balancing predictive performance with interpretability, consciously pairing necessary "black box" models with domain-specific XAI frameworks [1, 7].
- **Human-Computer Interaction (HCI):** Defining the target user (e.g., engineer vs. manager) and tailoring the explanation's interface and modality to their cognitive load [1].
- **Cybersecurity-by-Design:** Integrating principles from frameworks like the Explainable Resiliency Graph (ERG) to model how the system will explain cyber-physical attacks [1, 8].

5.2 Phase 2: Human-Centered Evaluation (HCE)

Addressing the "Evaluation Gap," this phase dictates that an explanation's quality cannot be measured purely by algorithmic fidelity; it must be evaluated via formal user studies to measure its operational utility [1, 15]. We synthesize the following core metrics as the standard for evaluation:

- **Trust & Satisfaction:** Measuring the user's confidence in the system's recommendations and the subjective usability of the interface [16, 17].
- **Objective Understanding:** Verifying the user's mental model via proxy tasks (e.g., predicting the AI's behavior in novel scenarios) [18].
- **Actionability & Reasonableness:** Assessing the user's ability to execute a correct physical response based on the explanation, and evaluating how coherently the AI's reasoning aligns with established engineering physics [6].
- **Query-Based Interaction:** Facilitating dynamic "what-if" explorations to fulfill specific human inquiries [19].

5.3 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.

5.3.1 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.

5.3.2 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.

5.3.3 The "Glass Box" Interface

A human-machine interface was constructed using **Streamlit** to serve as the central dashboard for the CPS facility manager. Figure 4 displays the real-time sensor telemetry where anomalies are actively tracked.

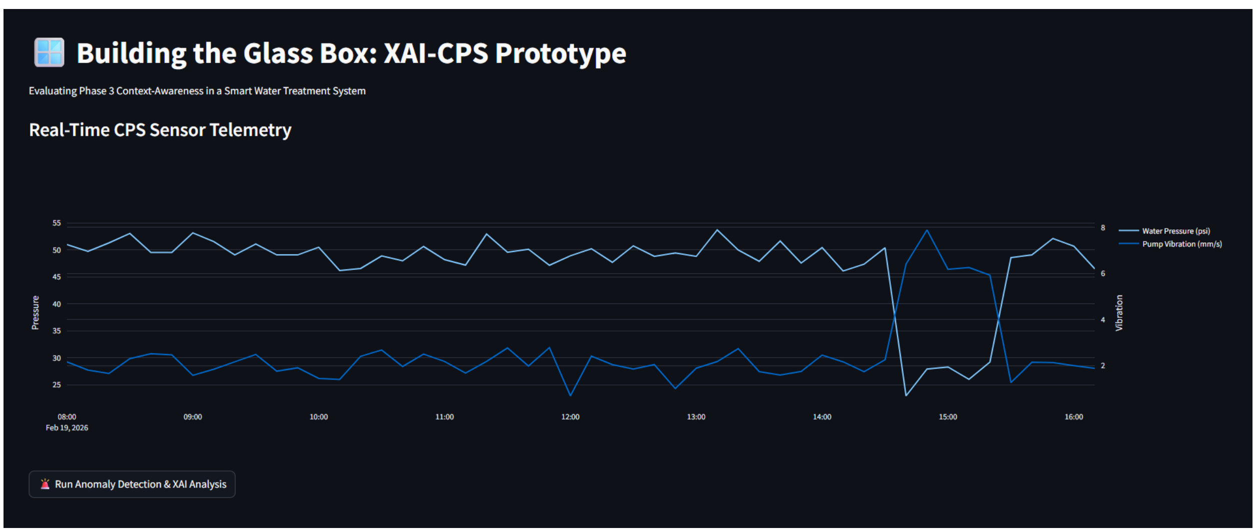

Figure 4: The "Glass Box" Streamlit dashboard displaying real-time CPS sensor telemetry and the injected anomaly window.

When the anomaly is triggered, the dashboard explicitly contrasts the two explanation paradigms side-by-side, as shown in Figure 5:

- **System A (Traditional XAI):** Provides a context-agnostic diagnosis, incorrectly suggesting a critical mechanical 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.

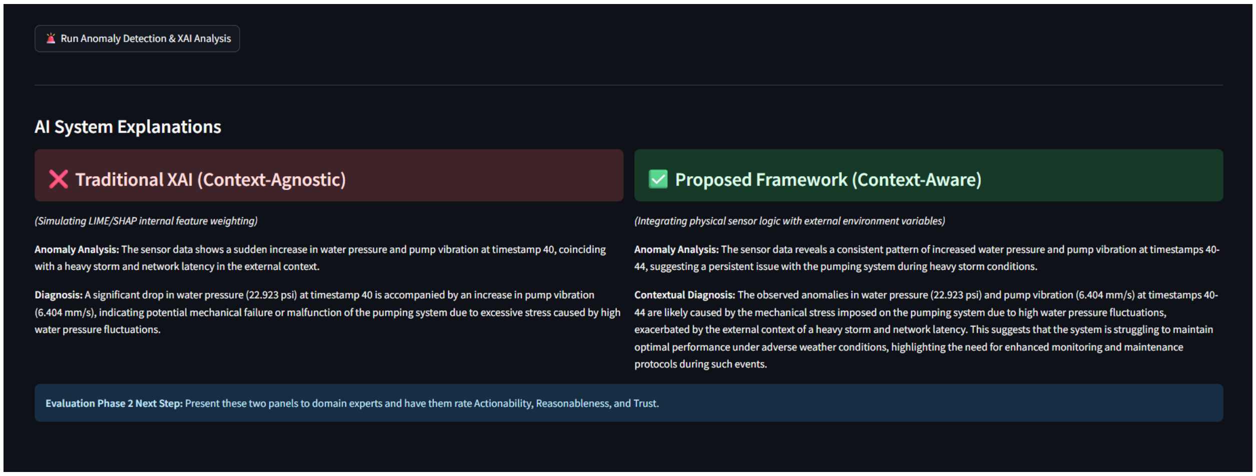

Figure 5: Side-by-side empirical comparison generated by the multi-agent system, contrasting traditional context-agnostic XAI with the proposed context-aware framework.

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.