Toward Trustworthy Diagnostic Reasoning for AI-Based Fault Diagnosis in Safety-Critical Systems

Ruixue Li, Katrina Groth
Center for Risk and Reliability, University of Maryland, College Park, MD, US, {ruilia21, kgroth}@umd.edu

Abstract: AI-based fault diagnosis is becoming increasingly critical in safety-critical industrial systems. However, predictive performance alone is insufficient to support trustworthy decision-making. A diagnostic model may successfully detect abnormal conditions while still failing to differentiate root causes from symptoms, capture credible fault propagation mechanisms, or generate explanations that operators can reliably interpret and act on. This perspective paper argues that the next challenge for AI-assisted fault diagnosis lies in trustworthy diagnostic reasoning, reasoning that is causally grounded, physically meaningful, and aligned with human decision-making requirements.

Recent advances in deep learning have significantly enhanced the representational capabilities of diagnostic AI, yet they simultaneously introduce new challenges related to trustworthiness. In safety-critical environments, weakly grounded graph construction or plausible-looking explanations may amplify spurious correlations and produce diagnoses that are correct in prediction while remaining difficult to trust in reasoning.

This paper positions knowledge-informed graph reasoning as a promising direction toward more trustworthy fault diagnosis. Industrial systems already embed substantial structured knowledge, including physical topology, control logic, material and energy flows, functional dependencies, safety constraints, and operator procedures. When incorporated appropriately, such knowledge can help constrain graph relationships and support explanations that better reflect physical propagation and operator reasoning. Functional reasoning paradigms inspired by Multilevel Flow Modeling (MFM) are discussed conceptually as one potential pathway for connecting learned diagnostic evidence with physical and functional system semantics.

1. INTRODUCTION

In safety-critical industrial systems, fault diagnosis cannot be treated solely as a problem of anomaly detection, fault classification, or predictive accuracy alone. While these capabilities are essential, they do not by themselves verify whether an AI system has correctly identified the root cause of an event, reconstructed a credible fault propagation pathway, or produced information that operators and Probabilistic Risk Assessment (PRA) analysts can meaningfully use for decision-making. Existing literature on industrial Fault Detection and Diagnosis (FDD) consistently highlights weak root-cause reasoning, limited interpretability, fragile graph structures, data scarcity, and deployment uncertainties as persistent limitations within the field [1], [2]. Consequently, for safety-critical applications, the central question is not merely whether a model can predict the correct fault category, but rather whether its diagnostic reasoning process is physically meaningful, causally sound, and practically valuable at the operational level.

Recent AI methodologies have significantly expanded the representational capabilities within the field of fault diagnosis. Deep learning enables the extraction of complex, non-linear patterns from industrial signals; Graph Neural Networks (GNNs) facilitate the representation of a system's structural interconnections; and knowledge-based or causality-driven AI methods allow for the integration of prior engineering semantics into the learning process [3], [4]. However, these technological advancements simultaneously introduce new challenges regarding trustworthiness. The construction of graph structures may encode statistical similarities, physical topologies, functional dependencies, or temporal causalities. Yet, the diagnostic conclusions supported by these distinct types of evidence may be fundamentally different. Similarly, visualization outputs, such as attention weights, saliency maps, and

graph activation values, may appear interpretable on the surface, but they do not necessarily constitute a factually valid causal explanation [5]. Consequently, strong performance on benchmark datasets may still coexist with weak root-cause localization, unstable explanations, or physically misleading reasoning.

This paper argues that the evaluation of safety-critical diagnostic AI systems should treat them as possessing a diagnostic reasoning capability, rather than confining the assessment solely to the scope of classification tasks. This paper proposes a novel perspective on trustworthy graph-based diagnostics, establishing graph structure construction, the validity of explanatory results, causal consistency, and knowledge-based reasoning capabilities as the core points of this field. The study explores function-based reasoning methods, inspired by Multi-Level Flow Modeling (MFM), as a high-level conceptual approach for realizing physically meaningful graph-based diagnostics. The remainder of this paper traces the methodological trajectory of fault diagnosis techniques as they evolve toward graph-based and knowledge-based paradigms; examines the construction of graph structures as a critical issue linked to trustworthiness; discusses the limitations inherent in evaluation paradigms centered solely on accuracy; and outlines future directions for the development of diagnostic AI technologies grounded in causal logic and closely aligned with operator requirements.

2. EVOLUTION OF AI-BASED FAULT DIAGNOSIS

The development of AI-based fault diagnosis can be understood as a gradual expansion in what diagnostic models are expected to represent. Classical data-driven FDD methods established the basic tasks of detecting abnormal operation, classifying known fault types, and, in limited cases, supporting isolation or root-cause analysis. Deep learning then shifted attention from hand-crafted or projection-based features toward learned representations of complex industrial signals [6]. Graph-based methods extended this trajectory by treating industrial systems as networks of interacting variables, sensors, components, or states. More recent knowledge-informed and causal approaches respond to the limitations of purely data-driven structure by asking whether diagnostic reasoning can be made physically meaningful, causally plausible, and useful to operators in safety-critical settings [7].

This evolution should not be read as a simple replacement of older methods by newer ones. Rather, each methodological stage addresses one set of limitations while exposing another. Classical FDD made the industrial diagnosis problem operationally concrete, but remained constrained by application specificity, limited root-cause reasoning, and weak handling of multiple or simultaneous faults [6]. Deep learning improved feature learning for nonlinear, high-dimensional, non-Gaussian, multimodal, time-varying, and autocorrelated process data, but did not eliminate data scarcity, distribution shift, interpretability, or deployment constraints [8], [9]. GNNs made relational structure explicit, but introduced graph construction as a safety-relevant modeling assumption rather than a neutral preprocessing step [3], [10]. Knowledge-informed and causal AI now point toward diagnostic reasoning that is not only predictive, but also physically and causally accountable [7], [11], [12].

2.1 Classical And Data-Driven Fault Detection And Diagnosis

Classical and data-driven FDD methods provide the baseline vocabulary for contemporary AI-assisted diagnosis. Chemical-process FDD reviews describe supervised and unsupervised methods such as Principal Component Analysis (PCA), Independent Component Analysis (ICA), Partial Least Squares (PLS), clustering, Bayesian networks, random forests, support vector machines, neural networks, and hybrid approaches [6]. Their importance is not only methodological; they formalized the distinction between detecting abnormal operation, assigning a fault class, and supporting diagnosis or isolation. That distinction remains central for safety-critical systems, because a correct fault label is not necessarily a sufficient basis for intervention.

The major limitations of classical FDD also remain relevant. The literature identifies application-specific performance, costly or imbalanced labels, and a frequent emphasis on detection or single-fault classification rather than root-cause diagnosis or multiple simultaneous faults [6]. These weaknesses

motivate later AI methods, but they also define the trustworthiness problem those methods inherit: safety-critical diagnosis must explain why an abnormal condition occurred, whether the identified fault is a source or symptom, and how the condition may propagate.

2.2 Deep Representation Learning For Industrial Diagnosis

Deep representation learning expanded FDD by reducing reliance on manually engineered features. Reviews of process FDD and industrial diagnosis/prognosis identify Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), Autoencoders, Generative Adversarial Networks (GANs), Transformers, and GNNs as major deep-learning families for diagnosis and prognosis tasks [4], [8]. This shift is well motivated because industrial process data may be nonlinear, high-dimensional, non-Gaussian, autocorrelated, multimodal, and time-varying [3]. Deep models can therefore learn latent temporal or signal patterns that are difficult to encode manually [4], [8].

However, deep learning has not entirely eliminated the challenges inherent to the FDD domain. Limited fault data become rare-event and imbalanced-learning problems; changing operation becomes distribution shift; interacting faults become compound or multi-label diagnosis; and deployment raises questions of computational cost, robustness, calibration, and maintainability [4], [8]. Interpretability also remains unresolved. Attention and Transformer-style methods can emphasize informative time steps, signal regions, channels, or features, and attention-based intelligent fault diagnosis has been organized into recurrent-based, convolution-based, and self-attention-based approaches [5]. However, attention weights are not automatically causal or physically meaningful explanations, and the attention literature itself identifies unresolved issues in interpretability, robustness, small-sample learning, computational cost, and integration with domain mechanisms [5]. For a trustworthy AI perspective, the key lesson is that merely enhancing the capabilities of representation learning does not automatically validate the efficacy of diagnostic outcomes, nor does it guarantee that the model's reasoning logic aligns with the cognitive patterns of human operators.

2.3 Graph-Based Fault Diagnosis And Prognostics

Graph-based fault diagnosis responds to the relational structure of industrial systems. Faults propagate through interacting components, sensors, process variables, control loops, material flows, and energy-transfer paths. GNN reviews describe graph representations as useful when diagnostic data contain relationships among samples, variables, sensors, or components, while Prognostic and Health Management (PHM) benchmarks treat graph construction, graph convolution, pooling, node-level diagnosis, graph-level diagnosis, and prognosis as central design dimensions [3], [13]. System-level PHM work likewise shows how components, sensors, and interactions can be represented as graph structures for health-state assessment [10].

The conceptual advance is a move from flat feature vectors toward relationship-centered diagnosis. But this advance creates a new trustworthiness problem: the graph itself encodes assumptions about what relationships matter. k-Nearest Neighbor (k-NN) or similarity graphs, prior-knowledge graphs, matrix-completion approaches, learned association graphs, system-interaction graphs, propagation-path graphs, functional models, Bayesian networks graphs, and temporal causal graphs support different diagnostic claims [3], [10], [13], [14], [15]. If edges encode weak, noisy, stale, or spurious relationships, message passing may amplify misleading structural features and produce explanations that appear plausible without being physically meaningful [3].

Recent research findings in graph-based fault diagnosis highlight both the potential of this field and the aspects that still need careful consideration. Propagation-path neural FDD embeds known chemical-process fault propagation paths into graph convolutional layers, but its evaluation is limited to selected Tennessee Eastman Process faults and leaves unknown-fault and multimodal or non-stationary settings unresolved [14]. Temporal causal graph diagnosis for nuclear power plants integrates expert priors with data and reports strong simulated results, but still identifies limitations in root-cause and propagation-path localization [15]. These examples support the transition to later sections of this paper: graph-based

diagnosis is promising precisely because it can represent system structure, but it becomes trustworthy only when graph construction, causal plausibility, and explanation validity are themselves evaluated.

2.4 Knowledge-Informed And Causal AI

Knowledge-informed and causal AI have emerged partly in response to the black-box limitations of deep learning. Informed machine learning classifies prior knowledge by source, representation, and integration point, including equations, simulations, logic rules, knowledge graphs, constraints, and human feedback [11]. Systems contain engineering knowledge about topology, physical flows, functional dependencies, control logic, failure modes, and operational constraints [11], [12]. Domain-knowledge integration reviews similarly describe input-level, loss-level, and architecture-level integration strategies, while emphasizing barriers such as differentiability, scalability, cyclic dependencies, and optimization difficulty [12].

Symbolic, functional, and causal AI approaches demonstrate how such domain knowledge can facilitate deeper diagnostic reasoning; however, this does not imply that the associated challenges have been fully resolved. Currently, researchers have successfully integrated symbolic domain knowledge into GNN through techniques such as node augmentation and the incorporation of relationships derived via Integer Linear Programming (ILP)[12]. Research based on Multi-Functional Modeling (MFM) indicates that functional knowledge can effectively support the visualization of diagnostic processes, aligned with the logic of physical flows, and aid in safety reasoning that combines both symbolic and probabilistic methods; concurrently, these studies also reveal lingering limitations regarding usability assessment, operational mode coverage, and the completeness of validation data [16]. Benchmark results for causality-enhanced Graph Neural Networks (GNNs) indicate that, in certain graph classification scenarios, the introduction of causal components indeed contributes to improving a model's robustness or generalization capabilities. However, current empirical results remain largely constrained by the specific characteristics of the datasets employed and have not yet reached the level of maturity and validation required to support industrial-scale FDD applications [7].

Furthermore, research focusing on time-series causal graph diagnostics in industrial settings, as well as FDD methods based on fault propagation paths, further highlights that precisely localizing the root causes of faults and achieving the online, dynamic discovery of fault propagation paths remain critical open challenges that urgently require resolution [14], [15].

It becomes evident that establishing a trustworthy diagnostic reasoning framework is becoming increasingly important. The core question currently demanding an answer is no longer confined to whether artificial intelligence can accurately classify faults, but rather whether it can underpin a higher-level diagnostic paradigm, one in which the constructed graph structures possess logical and practical significance; the provided diagnostic explanations remain faithful to the raw data while offering clear physical interpretability; and the proposed causal assertions clearly distinguish the root causes of faults from their superficial symptoms [3], [7], [16]. This transition sets the stage for the discussions in the subsequent chapters: graph construction must be treated as a matter of trustworthiness; diagnostic validity should not be confined solely to the assessment of accuracy; and knowledge-based reasoning, unless rigorously validated through established methodologies, must remain at a conceptual level, firmly grounded in empirical evidence.

3. GRAPH CONSTRUCTION AS A TRUSTWORTHINESS PROBLEM

Graph-based diagnosis is attractive for safety-critical industrial systems because faults rarely occur as isolated feature deviations. They propagate through components, sensors, control loops, process variables, material and energy flows, and functional dependencies. GNN-based FDD and PHM studies therefore, motivate a shift from flat feature vectors toward relational representations in which nodes and edges encode diagnostically relevant structure [3]. From a trustworthy-AI perspective, however,

the crucial issue is not only whether a graph is used, but what kind of relationship the graph claims to represent.

Graph construction is a modeling assumption and a relational inductive bias. A graph defines which entities can exchange information, which dependencies are amplified through message passing, and which parts of the system may appear important in a subsequent explanation. Existing graph-based diagnosis work uses or motivates several kinds of graph evidence, including statistical similarity, physical topology, process interaction, fault propagation pathways, functional dependencies, temporal causal dependencies, and Bayesian or PRA-related relations [3], [13], [14], [15], [17]. As a result, different graph structures support different diagnostic claims and forms of reasoning. Table 1 summarizes representative graph evidence types commonly used in graph-based industrial diagnosis.

Table 1: Representative Graph Evidence Types and Their Diagnostic Implications

Graph type	Encoded relationship	Diagnostic strength	Main limitation
Similarity graph/ k-NN graph	Statistical correlation	Captures correlated patterns	May encode spurious relationships
Process interaction graph	Operational dependency	Operational dependency	Sensitive to operating conditions
Topology graph	Physical connectivity	Reflects system architecture	Does not guarantee causal influence
Propagation-path graph	Hypothesized fault flow	Supports propagation reasoning	Limited by unseen faults
Functional / MFM graph	Functional and means-end dependency	Physically meaningful interpretation	Requires expert knowledge and formalization
Temporal causal graph	Lagged or instantaneous causal dependency	Supports root-cause localization	Sensitive to assumptions and data quality
Bayesian / PRA relation graph	Probabilistic dependency	Supports risk-informed reasoning	Depends on prior-model validity

The comparison highlights that graph construction is not merely a computational design choice, but a modeling assumption about what relationships are considered diagnostically meaningful. A graph that improves predictive performance may still fail to represent physically meaningful propagation pathways or causally relevant dependencies. The efficacy of graph-based diagnostic methods fundamentally hinges on how the graph itself is constructed. Unlike traditional deep learning models, where relationships between variables are often implicitly learned from data, GNNs require that the relationships between nodes and edges be explicitly specified or learned. Consequently, graph construction is no longer merely a neutral pre-processing step; it evolves into a modeling assumption that directly shapes the process of diagnostic inference [3]. The graph structure dictates which components exchange information, which dependencies are emphasized, and which propagation paths are implicitly treated as physically meaningful.

Graph construction should be evaluated as part of trustworthy diagnostic reasoning. The question is not only whether graph-based AI improves classification, but whether its graph structure supports causally plausible, physically meaningful, and operator-interpretable reasoning. In this view, graph construction becomes inseparable from the trustworthiness of diagnostic reasoning itself. This moves the discussion naturally beyond accuracy toward causal consistency, explanation fidelity, and diagnostic validity.

4. BEYOND ACCURACY: CAUSAL CONSISTENCY AND EXPLANATION VALIDITY

The limitations of current evaluation practice can therefore be understood as a mismatch between predictive performance and trustworthy diagnostic reasoning. Existing benchmark protocols primarily

evaluate whether a model can correctly reproduce known fault labels under controlled test conditions. However, safety-critical diagnosis requires additional reasoning capabilities related to causality, propagation understanding, explanation validity, and operator decision support. Different evaluation perspectives, therefore, answer fundamentally different diagnostic questions, as summarized in Table 2.

Table 2: Distinguishing Prediction Accuracy, Causal Consistency, and Explanation Validity in Safety-Critical Diagnosis

Evaluation aspect	Central question	Limitation if used alone
Classification accuracy	Was the fault class predicted correctly?	Does not ensure root-cause understanding
Statistical dependency learning	Which variables are strongly associated?	Association does not imply causal propagation
Attention/saliency explanation	Which features or nodes influenced the prediction?	May produce plausible but unfaithful explanations
Explanation validity	Is the explanation faithful, physically meaningful, and operator-relevant?	Still insufficiently evaluated in current FDD literature
Causal consistency	Does the reasoning align with plausible fault mechanisms and propagation logic?	Difficult under distribution shift and unseen faults

The comparison illustrates that predictive success, causal validity, and explanation trustworthiness are related but non-equivalent objectives in safety-critical diagnosis. The resulting distinction is therefore between explanation plausibility and explanation validity: an explanation may look reasonable to an analyst, highlight familiar variables, or produce a visually coherent path while still being unfaithful to the model's computation, disconnected from physical propagation, or irrelevant to the operator's decision context

The distinction between predictive success and trustworthy reasoning is illustrated conceptually in Figure 1. Two diagnostic models may produce the same fault prediction while relying on fundamentally different underlying reasoning pathways. One pathway may align with physically meaningful propagation logic and causal system behavior, whereas the other may exploit shortcut correlations, operating-mode artifacts, or dataset-specific associations that do not represent valid fault mechanisms.

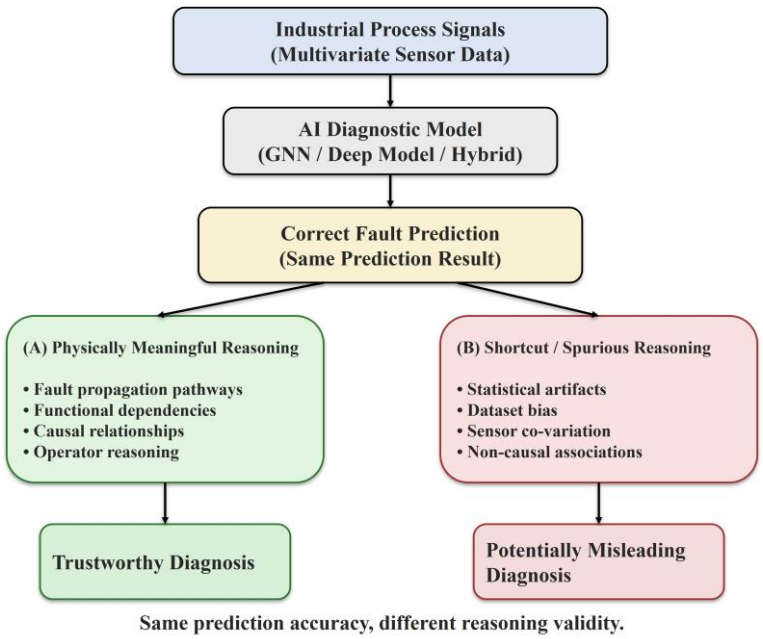

Figure 1: Two Diagnostic Reasoning Pathways Leading to the Same Prediction.

Figure 1 highlights that prediction accuracy and reasoning validity are related but non-equivalent objectives in safety-critical diagnosis. A model may achieve strong benchmark performance while relying on relationships that are physically weak, causally ambiguous, or operationally unstable. This distinction motivates the need for evaluation frameworks that assess not only predictive correctness, but also causal consistency, explanation fidelity, and operator-relevant interpretability.

5. KNOWLEDGE-INFORMED DIAGNOSTIC REASONING FOR SAFETY-CRITICAL SYSTEMS

The limitations discussed above point toward a broader research direction: safety-critical diagnostic AI should not reason from data alone when the system already contains structured engineering knowledge. Industrial plants encode topology, control logic, physical laws, flow relations, functional dependencies, safety constraints, maintenance knowledge, operating procedures, and expert judgment. Knowledge-informed AI is therefore attractive not only because fault data are scarce, but because diagnostic relationships need physical and functional meaning. The core challenge lies in how to integrate learning with engineering semantics, causal reasoning, and propagational perception interpretation, so that the learned model should be guided, examined, and interpreted through relationships with reasonable physical, functional, or causal significance.

5.1. Functional Knowledge Representations

Functional representations are highly valuable because they describe the objectives a system is intended to achieve, how underlying functions support higher-level goals [18]. This level of abstraction is crucial for fault diagnostics, as operators typically need to understand not only which variable has deviated, but also which specific function has been impaired, whether the deviation is a cause or merely a symptom, and how the disturbance propagates throughout the system. Consequently, means-end reasoning and physical-process reasoning provide a means of linking local observations to system-level consequences and operator-relevant explanations.

MFM is relevant in this context because functional representations can relate local observations to system-level goals, flows, and propagation behavior. Recent research involving MFM highlights the significance of such representations: MFM has been utilized to filter XAI explanations, ensuring that the diagnostic causes presented to operators align with the underlying physical processes; furthermore, MFM has been integrated with Bayesian inference to model functional fault propagation and assess process safety [16], [17]. It follows, therefore, that functional knowledge can provide a layer of semantic reasoning to determine whether diagnostic explanations align with the system's purpose, processes, and plausible propagation behaviors. It helps to reframe the diagnostic inquiry from "Which class of fault is predicted?" to "Which functions are affected? Which sources can account for these symptoms? And which consequences are critical for a safe response?"

5.2. Toward Physically Meaningful Graph-Based Diagnosis

Graph-based diagnosis becomes more trustworthy when graph relationships carry interpretable engineering meaning. Knowledge-guided graph reasoning should therefore ask what each relationship represents, what evidence supports it, and whether inference over that relationship is causally and physically plausible.

Physically meaningful graph-based diagnosis requires reasoning structures that connect observed industrial behavior, engineering knowledge, and operator-relevant interpretation. The relationship among these layers is illustrated conceptually in Figure 2.

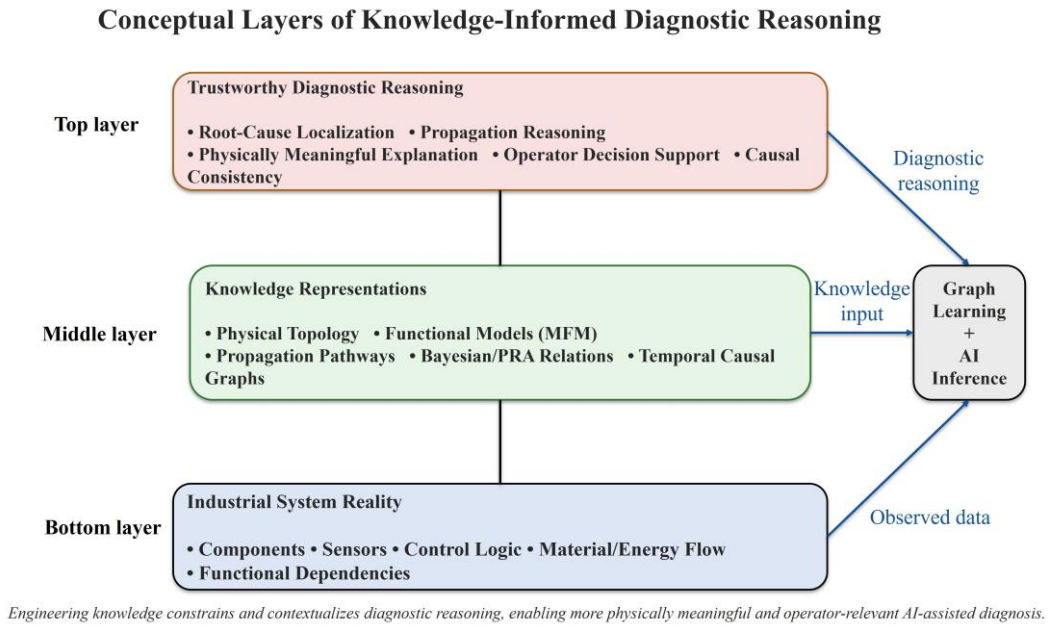

Figure 2: Conceptual Layers of Knowledge-Informed Diagnostic Reasoning for Safety-Critical Systems.

Figure 2 emphasizes that trustworthy diagnostic reasoning depends not only on predictive learning capability, but also on how engineering knowledge constrains, contextualizes, and interprets diagnostic relationships.

This perspective motivates hybrid symbolic and graph-learning paradigms for safety-critical diagnosis. Functional knowledge can represent system entities and propagation relationships at a conceptual level, while graph learning can support relational inference over complex process data. Causal reasoning may further help distinguish initiating causes from propagated symptoms and improve propagation-aware interpretation.

6. CONCLUSION

AI-based fault diagnosis is evolving from a data-driven classification paradigm into a diagnostic reasoning paradigm grounded in graph structures, integrating domain knowledge, and centered on causality. This evolution reflects the reality of safety-critical industrial systems: in such systems, faults do not arise from isolated feature deviations, but rather from the interplay among component interactions, material and energy flows, control actions, functional dependencies, and fault propagation paths. However, the literature reviewed in this paper also indicates that merely enhancing a model's representational capacity does not automatically guarantee the reliability of diagnostic results. Graph structures may encode misleading associations; attention mechanisms or salient features may be misconstrued as genuine causal explanations; and exceptionally high accuracy rates on benchmarks often coexist with deficiencies such as weak root-cause localization capabilities, limited causal validity, and low practical utility for human operators.

Consequently, the central argument of this paper is that, for safety-critical diagnostic AI, predictive accuracy alone is insufficient. Future diagnostic systems must also be evaluated based on their causal consistency, the fidelity of their explanations, and the plausibility of their physical interpretations. These criteria determine whether an AI-assisted diagnostic system can provide sufficiently compelling and defensible support for the reasoning processes concerning fault origins, symptom manifestations, propagation paths, consequential impacts, and mitigation strategies.

Graph-based reasoning techniques that integrate domain knowledge represent a promising pathway toward achieving this objective. By incorporating methods such as functional representations,

knowledge of fault propagation paths, causal graph models, symbolic relationships, and probabilistic safety reasoning, these techniques help bridge the gap between the patterns learned by a model and the engineering semantics and physically meaningful explanations required in practice.

For safety-critical industrial systems and decision support applications, the challenge extends beyond merely determining whether an AI can detect a fault; it fundamentally concerns whether the AI's reasoning process is trustworthy within the context of actual operational decision-making environments. Therefore, the next frontier in the field of safety-critical diagnostic AI is establishing a capability for trustworthy reasoning, one that is grounded in physical functional principles, the causal mechanisms of fault propagation, and the specific decision-making requirements of human operators.

References

[1] Y. Xu, Y. Sun, J. Wan, X. Liu, and Z. Song, "Industrial Big Data for Fault Diagnosis: Taxonomy, Review, and Applications," *IEEE Access*, vol. 5, pp. 17368–17380, 2017, doi: 10.1109/ACCESS.2017.2731945.

[2] S. Dash and V. Venkatasubramanian, "Challenges in the industrial applications of fault diagnostic systems," *Comput. Chem. Eng.*, vol. 24, no. 2, pp. 785–791, Jul. 2000, doi: 10.1016/S0098-1354(00)00374-4.

[3] Z. Chen *et al.*, "Graph neural network-based fault diagnosis: a review," Nov. 16, 2021, *arXiv*: arXiv:2111.08185. doi: 10.48550/arXiv.2111.08185.

[4] S. Qiu *et al.*, "Deep Learning Techniques in Intelligent Fault Diagnosis and Prognosis for Industrial Systems: A Review," *Sensors*, vol. 23, no. 3, p. 1305, Jan. 2023, doi: 10.3390/s23031305.

[5] H. Lv, J. Chen, T. Pan, T. Zhang, Y. Feng, and S. Liu, "Attention mechanism in intelligent fault diagnosis of machinery: A review of technique and application," *Measurement*, vol. 199, p. 111594, Aug. 2022, doi: 10.1016/j.measurement.2022.111594.

[6] S. A. A. Taqvi, H. Zabiri, L. D. Tufa, F. Uddin, S. A. Fatima, and A. S. Maulud, "A Review on Data-Driven Learning Approaches for Fault Detection and Diagnosis in Chemical Processes," *ChemBioEng Rev.*, vol. 8, no. 3, pp. 239–259, Jun. 2021, doi: 10.1002/cben.202000027.

[7] S. Job *et al.*, "Causal integration in graph neural networks toward enhanced classification: benchmarking and advancements for robust performance," *World Wide Web*, vol. 28, no. 3, p. 30, May 2025, doi: 10.1007/s11280-025-01343-1.

[8] J. Yu and Y. Zhang, "Challenges and opportunities of deep learning-based process fault detection and diagnosis: a review," *Neural Comput. Appl.*, vol. 35, no. 1, pp. 211–252, Jan. 2023, doi: 10.1007/s00521-022-08017-3.

[9] K. Lagemann, C. Lagemann, B. Taschler, and S. Mukherjee, "Deep learning of causal structures in high dimensions under data limitations," *Nat. Mach. Intell.*, vol. 5, no. 11, pp. 1306–1316, Nov. 2023, doi: 10.1038/s42256-023-00744-z.

[10] A. R.-T. Palazuelos, E. L. Droguett, and K. M. Groth, "A System-Level Prognostics and Health Management Framework based on Graph Convolutional Neural Networks," in *Proceedings of the 30th European Safety and Reliability Conference and 15th Probabilistic Safety Assessment and Management Conference*, Research Publishing Services, 2020, pp. 1688–1694. doi: 10.3850/978-981-14-8593-0_4165-cd.

[11] L. Von Rueden *et al.*, "Informed Machine Learning - A Taxonomy and Survey of Integrating Prior Knowledge into Learning Systems," *IEEE Trans. Knowl. Data Eng.*, pp. 1–1, 2021, doi: 10.1109/TKDE.2021.3079836.

[12] T. Dash, A. Srinivasan, and L. Vig, "Incorporating symbolic domain knowledge into graph neural networks," *Mach. Learn.*, vol. 110, no. 7, pp. 1609–1636, Jul. 2021, doi: 10.1007/s10994-021-05966-z.

[13] T. Li, Z. Zhou, S. Li, C. Sun, R. Yan, and X. Chen, "The emerging graph neural networks for intelligent fault diagnostics and prognostics: A guideline and a benchmark study," *Mech. Syst. Signal Process.*, vol. 168, p. 108653, Apr. 2022, doi: 10.1016/j.ymssp.2021.108653.

[14] B. Nguyen and M. Chioua, "A propagation path-based interpretable neural network model for fault detection and diagnosis in chemical process systems," *Control Eng. Pract.*, vol. 150, p. 105988, Sep. 2024, doi: 10.1016/j.conengprac.2024.105988.
[15] J. Liu, K. Pan, E. Zio, and Y. Yao, "Temporal causal graph-based attention gated recurrent unit for interpretable fault diagnosis in nuclear power plants," *Reliab. Eng. Syst. Saf.*, vol. 273, p. 112332, Sep. 2026, doi: 10.1016/j.ress.2026.112332.
[16] J. H. Shin, J. S. Kang, J. M. Kim, and S. J. Lee, "Concept of understandable diagnostic cause visualization with explainable AI and multilevel flow modeling," *Nucl. Eng. Technol.*, vol. 57, no. 8, p. 103589, Aug. 2025, doi: 10.1016/j.net.2025.103589.
[17] R. Li, J. Wu, O. Ravn, and X. Zhang, "An integrated framework for functional model-based safety assessment of process systems using Cloud-Bayesian network," *Reliab. Eng. Syst. Saf.*, vol. 271, p. 112231, Jul. 2026, doi: 10.1016/j.ress.2026.112231.
[18] M. Lind, *Foundations for Functional Modeling of Technical Artefacts*. in Design Research Foundations. Cham: Springer International Publishing, 2024. Accessed: Feb. 02, 2026. [Online]. Available: https://link.springer.com/10.1007/978-3-031-45918-4