Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

An Explainable Artificial Intelligence Framework for Decision-Making in Critical Information Technology Systems

SAI DOONDI KOTHAPALLI

saidoondi7@gmail.com

Atlantis University https://orcid.org/0009-0005-6580-1268

Research Article

Keywords: Explainable artificial intelligence, Critical IT systems, Decision support, SHAP, LIME, Intrusion detection, AI governance

Posted Date: August 14th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10682801/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

An Explainable Artificial Intelligence Framework for Decision-Making in Critical Information Technology Systems

Sai Doondi Kothapalli [1]

Abstract

AI systems are now being deployed inside the critical IT systems like intrusion detection systems, IT operations analytics (AIOps), cloud governance platforms, clinical decision support systems and more. Artificial intelligence (AI) systems are now being incorporated into critical information technology (IT) infrastructure such as intrusion detection systems, IT operations analytics (AIOps), cloud governance platforms, clinical decision support systems and more. Deep learning and ensemble methods provide better predictive accuracy but have the disadvantage of being "black-box" algorithms, which makes them less trusted by operators, harder to regulate and create the potential for undetected propagation of errors in critical applications. Explainable AI (XAI) is now the main solution to this lack of transparency, though there is a disciplinary divide in the literature and no synthesis of technical methods of explainability and operational, human factors and governance needs of critical IT systems exists. This review aims to identify, integrate and apply 32 peer-reviewed and archival research papers published mainly during the period of 2019 to 2025, in addition to foundational research, to build an integrated XAI framework for critical IT decision making. The methodology is guided by a structured narrative-systematic review protocol covering explainability techniques, applications in the domain of cybersecurity, AIOps, and infrastructure in the healthcare domain, and governance instruments like the NIST AI Risk Management Framework. Post-hoc model-agnostic methods, especially SHAP and LIME, are the most applied methods today, especially in intrusion detection and financial risk applications, and are significantly preferred over intrinsically interpretable and attention-based methods, despite the latter's better transparency-performance trade-offs, according to the review. A conceptual five-layer model is suggested, that combines the concepts of data provenance, the generation of explainability, interpretation with human in the loop, and the governing process that is auditable. Persistent issues are highlighted such as the instability of explanations, computational complexity, and inconsistent evaluation of explanations, as well as the lack of a standardised fidelity measure. The review shows that instead of single explainability techniques, it is a need for coordinated progress in technical methods, how to design interfaces, and the alignment of the regulations to make operational explainability in critical systems possible.

Keywords: *Explainable artificial intelligence; Critical IT systems; Decision support; SHAP; LIME; Intrusion detection; AI governance.*

1. Introduction

Advancements in artificial intelligence (AI) in information technology (IT) infrastructures have changed the way organisations identify threats, allocate computing resources and aid in making operator decisions. Deep learning and ensemble machine learning models are increasingly used to process high volume, high velocity data streams in critical systems—those whose failure or compromise have significant operational, financial or safety implications such

as network intrusion detection systems, cloud-scale IT operations platforms, financial transaction monitoring, or clinical decision support. Many critical systems—those with potentially significant operational, financial or safety implications when they fail or are compromised—are increasingly dependent on deep learning and ensemble machine learning models for processing high volume, high velocity data streams, such as network intrusion detection systems, cloud-scale IT operations platforms, financial transaction monitoring, or clinical decision support [1,2]. Such models are often more accurate than conventional rule-based systems, but cannot be easily understood by the human user who ultimately has to approve, reject, or validate their results.

The opacity of this 'black-box' issue is not an incidental annoyance but a problem inherent in the safe use of AI in critical environments. If an intrusion detection system incorrectly identifies network traffic as malicious, a security analyst can't easily pick out a real attack from a false positive, and a verdict in a subsequent audit or trial cannot be substantiated. [3,4] Likewise, if the AIOps platform suggests an automated remediation measure in the cloud, engineers need to be assured that they are receiving a recommendation that is a true causal signal, and not some coincidence in the training data. Explainable artificial intelligence (XAI) has thus made its way to a new research agenda to make the intrinsic mechanisms of complex models understandable to humans without sacrificing predictive accuracy [3].

The Defense Advanced Research Projects Agency (DARPA) XAI programme, which was started in 2016, was one of the first coordinated efforts to formalise this objective and considers explainability as a requirement for proper trust calibration between human operators and autonomous systems [1,2]. The field has since grown in scale, with a variety of post-hoc explanation techniques now available (including Local Interpretable Model-agnostic Explanations (LIME) [6] and Shapley Additive explanations (SHAP) [7]) together with intrinsically interpretable model classes, and a burgeoning body of human-computer interaction research aimed at understanding how explanations are actually consumed by end users [9,22].

While the field of XAI has seen much growth, the use of XAI for critical IT systems is comparatively underdeveloped, and different from the more widely explored areas of medical imaging and consumer-focused recommender systems. Current surveys mostly focus on either the technical aspects of explainability methods [3,5] or the application of explainability methods in specific contexts, such as intrusion detection [17,18,19] or AIOps [32], without connecting them to a practical framework that can be adopted in operational contexts of high importance for IT systems. This fragmentation makes it difficult for practitioners to determine which techniques to use for which types of problem, to ensure explanations are correct, and to understand that explainability can play a role in other governance concerns, like those from the National Institute of Standards and Technology's AI Risk Management Framework [29].

To fill this gap, this is a review that aims to synthesize recent literature into an integrated framework to explain decision-making in critical IT systems. The specific goals are threefold: 1) describe, in detail, the main explainability techniques that are currently applied in critical IT contexts, and their relative advantages and disadvantages; 2) review how these techniques have been applied in representative critical IT domains, such as network security, IT operations management and related to high-stakes decision-making in other domains; 3) suggest a consolidated conceptual framework linking the processes of technical explanation generation to human interpretation and governance processes. The rest of this article is organized as follows. The review method is given in Section 2. In Section 3, the literature is

reviewed from the perspective of the explainability technique and application domain. In Section 4, the problem statement motivating the proposed framework is stated. The results of the synthesis, including the proposed framework and comparative tables, are presented in section 5. The implications of these findings, challenges and limitations are discussed in Section 6 and conclusions are presented in Section 8.

2. Methodology

The designed narrative-systematic method is a structured approach that involves a traditional review and elements of a protocol-driven source identification as in the case of a systematic review, suitable for a technically heterogeneous and rapidly evolving field like XAI [4]. Due to methodological heterogeneity of primary studies (controlled technical benchmarking, qualitative human-computer interaction studies and policy analysis), a completely quantitative systematic review with meta-analytic pooling was deemed unsuitable.

Structured searches were conducted in academic databases and preprint repositories, such as IEEE Xplore, ScienceDirect, ACM Digital Library, MDPI journals and arXiv, following the combinations of the following terms: explainable artificial intelligence, XAI, interpretable machine learning, critical infrastructure, intrusion detection, AIOps, decision support, and trustworthy AI. The search was limited to works published from 2019 to 2025, but some important works predating this period were also included, as they set the foundations of concepts and techniques that are still central today, such as the DARPA XAI programme announcement [1], the LIME algorithm [6] and the SHAP framework [7].

The inclusion criteria were (i) the source needed to discuss explainability or interpretability of AI or machine learning models, (ii) it had to be published in a peer-reviewed journal, peer-reviewed conference proceedings, or a recognised technical report or technical standard, and (iii) it had to have direct relevance to a critical IT system, cybersecurity, IT operations, or closely related high-stakes decision domains that would inform the transferability of explainability methods to the context of IT. Exclusion criteria were applied to sources that did not contain methodological information relevant to explainability except in an unrelated consumer or entertainment context and sources without full text that allowed for verification of claims.

A total of 32 sources, after screening of titles, abstracts and full text, were selected from an initial pool identified in the search process for detailed synthesis. Extracted from each source were the explainability technique(s) used, the application field, strengths and limitations reported, and empirical results about explainability fidelity, computational requirements, and user trust. Extracted data were then thematically organized by three main aspects: (i) explainability technique (post-hoc model-agnostic, post-hoc model-specific, and intrinsically interpretable approaches), (ii) application domain within critical IT systems, and (iii) governance and human-factors considerations. The thematic organisation is a direct result of this literature review and the conceptual framework presented in the results section. The review is not a statistical meta-analysis, but is a narrative-synthetical one, following other recent reviews in the XAI literature that consider analogous technical fields of study in different problem areas [4,32].

3. Literature Review

The literature can be structured along three major threads: basic explainability concepts and methods in critical IT decision-making, use or implementation of explainability for critical IT systems applications, and explainability governance and human factors research for the evaluation, trust and regulation of explanations. All the strands are explored individually below.

3.1 Foundational Concepts and Explainability Techniques

The principles of XAI are grounded in a distinction found in the literature, and widely used, between transparency and "post-hoc" explainability [3,10]. Transparent models are those that reveal their decision-making processes explicitly, like linear regression, decision trees, and rule-based expert systems, while opaque models like large ensembles of deep neural networks and deep neural networks do not show their learning of the underlying mechanisms that guide their predictions [5,11].

In a high-stakes decision setting, practitioners should prefer inherently interpretable models over the efforts to explain the "black box" models, as long as they make predictions accurately (Rudin [11]). But there is a counter-argument that as stated above, with many unstructured or high-dimensional unstructured data types typical of network traffic, log files and sensor telemetry in critical IT systems, the intrinsically interpretable models often fail to outperform deep learning approaches [17,19].

LIME, proposed by Ribeiro, Singh and Guestrin [6] is among the post-hoc techniques, which work by corrupting a single instance in an input and fitting an interpretable model to the set of surrounding instances to obtain the local surrogate explanations. On the other hand cooperative game theory-based Shapley values [7] predict the output of an prediction using individual input features and have a number of consistency properties both in the local and global sense that are not guaranteed by LIME.

Across several evaluations, attributions generated with SHAP have proven to be more stable and theoretically sound than those with LIME, but are also much more costly to compute, especially when the feature space is large, as is often the case in network intrusion datasets [18,19]. To illustrate this sensitivity, Alvarez-Melis and Jaakkola [15] showed examples of several of the recent post-hoc interpretability techniques (such as LIME-style local approximations) being vulnerable to nigerent input changes. Among the different methods that have been emerged to handle the local and global mechanics of these feature interactions include partial dependence plots/accumulated local effects plots for better visualisation of the influence of a single feature or of all features collectively [14], localization mechanisms like in transformer and recurrent architectures that reveal the relative importance of input features and units [13] and counterfactual explanation techniques that explain the minimal alterations in the input needed to cause the model output to change.

Layer-wise relevance propagation and similar methods utilizing gradients provide efficient alternatives to other perturbation methods for convolutional and recurrent networks that are frequently used in log-based anomaly detection, as discussed by

Samek et al. [13]. Doshi-Velez and Kim [8] presented an influential framework for assessing the interpretability of a system, separating the three types of evaluation: application-grounded evaluation (expert user performing real tasks); human-grounded evaluation (with simulated versions of the original tasks); and functionally-grounded evaluation (without human subjects using the evaluation of model complexity as a proxy). This taxonomy is key to follow-up evaluation studies such as [27] which introduced a design and evaluation framework for interpretable machine learning systems that relates end user goals to various stages of systems design.

3.2 Explainable AI in Network Security and Intrusion Detection

In particular, practitioners have been most apalled by the move toward opaque machine learning models, creating an unusual amount of resistance to XAI in the security operations centre space, which also has high stakes for failure due to the consequences of intrusions being missed [17]. Sinha et al. [17] showed that feature attribution explanations helped security analysts extract information about the traffic's characteristics that lead to a certain alert classification from ensemble ID classifiers, thus minimizing the time required for analysts to verify results compared to those left without an explanation. Gaspar, Silva and Silva [18] first trained a multi-layer perceptron intrusion detection model for Internet of Things network traffic, and then attempted to falsify explanations generated by LIME and SHAP by selectively removing features found to be influential to validate the explanations.

They concluded that their results showed SHAP displayed higher inter-run consistency than LIME, but LIME produced results faster than SHAP for near real-time analyst workflows. In order to identify this lack of quantitative fidelity measurements, Arreche et al. [19] followed this tradition of comparing explanations extensively by systematically examining several black-box explainability methods on a number of intrusion detection datasets and suggesting quantitative fidelity measures to gauge the faithfulness of explanations of modeled behaviour. Kalutharage et al. [30] studied the use of XAI in IDS for an IoT environment, as the low computational power of an edge device prevalent in IoT deployments puts more constraints on explanation generation than those found in conventional enterprise network monitoring, prompting the use of lightweight generation of explanations rather than perturbation-based approaches. While classical sequential architectures have proven to be very successful at advancing the sophistication of novel and zero-day attack detection, the need for providing explanations post hoc has become more pressing with increasingly complex sequential architectures proposed by Hnamte et al. [20] using a two-stage deep architecture of autoencoders combined with long short-term memory networks for network intrusion detection.

A more recent advancement “embeds” LLM's in the flywheel of security operations explanations. Ali and Kostakos in [21] presented a system that integrates explainable AI, large language model based natural language explanations and machine learning-based anomaly detection, which increased the trust of the analysts and decreased the burden of interpretation when presented with raw feature attribution output, leading

them to believe that, while quantitative methods are here to stay, natural language explanation generation is poised to complement them.

3.3 Explainable AI in IT Operations and AIOps

Many Artificial Intelligence for IT Operations (AIOps) platforms leverage the analysis of logs, metrics and traces collected by the distributed IT infrastructure to automatically surface anomalies, investigate the root cause and predict failures at cloud scale [32]. While AIOps can be used independently, Diaz-De-Arcaya et al. [32] position AIOps in a larger context of other operational approaches such as DevOps, DataOps and MLOps; noting that the implicit nature of the machine learning models presents unique challenges to their operational trust because operators need to act quickly on system recommendations without the time to ensure their manual verification. The multifaceted nature of operational data makes explanations difficult, from structured metrics, to semi-structured log data and unstructured incident narratives, which each might need different approaches to explanation. Research papers on multi-modal data fusion for AIOps situations find that applying multiple feature attribution techniques and using cross-source correlation analysis allows explaining a root cause message based on multiple telemetry sources at once, a feature that is especially useful during a major incident response that involves joint causal chain identification for interdependent services.

The most common explainability techniques used in AIOps are still SHAP and LIME which are also the most common in network intrusion detection, while the use of rule-based post-hoc visualisation and log-pattern highlighting still has a high-level adoption because of its low information processing latency needed in operational environments. Finally, based on reviews of explainable AIOps for cloudscale DevOps automation by Nguyen et al., only a fraction of currently available commercial and research AIOps tools give some sort of post-hoc visualisations or additional chunks of logs; future systems should allow engineers to reject AI suggestions, and indicate why, to generate a feedback loop that can be used to improve the AI model and its explanation mechanism, respectively. This is consistent with some feedback and monitoring as part of this conceptual framework outlined in Section 5 of this review.

3.4 Explainable AI in Adjacent High-Stakes Decision Domains

This review's focus is on critical IT systems, but there are some canary-in-the-cooindustrial process control and clinical decision support systems in the adjacent vicinity from which methodologically transferable insights can be drawn – these are areas that have much more established evaluation standards for explainability than those related to cybersecurity and to IT operations. Tjoa et al. [25] offer a thorough survey of medical domain of XAI, which classifies the methods into two types: perceptive and interpretable structure; and urge to pay attention to domain-specific correctness criteria when designing explanations for medical use. Similarly, the Italian Society of Medical and Interventional Radiology (SIR)'s team of Neri et al. [24] state that, when creating new clinical workflows, explainability should be incorporated at

the beginning and should therefore not be considered as a post-production addition in the form of a visualisation layer. This brings explainability to the interface design of data and metrics in a security operations centre or a network operations centre. The concept of causability was introduced by Holzinger et al. [26] which they distinguish from the property of explainability; they believe that measures of causability should be used along with measures describing technical explainability in any system used for consequential human decision making. There are obvious parallels with the explanatory value of technically correct renditions of the internal workings of a model, when delivered without the proper contextualisation, which do not lend themselves to the analyst's causal prose about a security incident in critical contexts for IT. They found out that in safety-critical industrial control systems where there is a direct physical safety consequence, being able to interpret models is the preferable option over improving the marginal predictive capability of their intrinsic interpretability, in comparison to other variants of LX that they reviewed [23]. This stance was directly applicable to the trade-off between being interpretable and being predictive discussed in Section 5 of this review.

3.5 Human Factors, Trust, and Governance

One thread of the literature takes the simplistic first step in going beyond the technical problem of explanation generation and analyzes what human decision makers do with explanations when they receive them. To better inform their review of the subtleties of the problem of conveying explanations to end-users, Laato et al. [22] conducted a systematic review of the literature on end user explanations of AI systems, noting examples across a range of use cases. They found that advice on the content, form and timing of explanations needs to be personalized to match end user expertise, task context and that a single explanatory modality was not well suited to both experts and non-experts or end-users across the various scenarios. In contrast to the high dimensional mapping of contributions that are generally the output of SHAP under the umbrella of complex models, Miller argues [9] that in human explanation, it is key to apply a principle to good explanation: to consistently be contrastive, selective and socially situated. The findings by Vasconcelos et al. [31] suggest that not all explanations will necessarily enhance decision quality: in some experimental scenarios, providing explanations increased inappropriate behavior of relying on the AI recommendations, as users followed system outputs despite explanations being expected to lead the user to question the outputs. This discovery also has implications for critical IT operations: If a security alert or remediation hint is generated by an AI system the operator of a critical operation could face an operational risk by accepting it as they are essentially handing over a false sense of security. عَلَيْ explanation design must enable, but more importantly support, human operators to carry out critical evaluation, rather than assume it automatically happens. The National Institute of Standards and Technology's AI Risk Management Framework [29] says that explainability and interpretability should be considered key traits of systems that are trustworthy along with validity, reliability, safety, security, accountability, and fairness, and it urges organizations to record the nature of explainability of their

deployed models as part of their broader AI risk management processes. Organisations operating critical IT infrastructure in different jurisdictions will find it difficult to operate in the presence of these different regulatory regimes, in part due to the extent of variation in definition of adequate explanation across jurisdictions as reported by Nannini, Balayn and Smith [28]. In an early and often referenced study, Goodman and Flaxman [12] discuss the meaning of the provisions in the EU on data protection and their consequences for an ensuing "right to explanation" for automated decisions, that continues to be an issue in subsequent analyses, such as within the NIST framework and accompanying documents. Another important application of explainability that Mehrabi et al. [16] accelerate is the use of feature attribution methods as part of bias detection in critical IT systems, specifically, the ability to detect if protected or proxy features drive disproportionate decisions by automated systems, critical to identity and access management systems.Organisations operating critical IT infrastructure in different jurisdictions will find it difficult to operate in the presence of these different regulatory regimes, in part due to the extent of variation in definition of adequate explanation across jurisdictions as reported by Nannini, Balayn and Smith [28]. In an early and often referenced study, Goodman and Flaxman [12] discuss the meaning of the provisions in the EU on data protection and their consequences for an ensuing "right to explanation" for automated decisions, that continues to be an issue in subsequent analyses, such as within the NIST framework and accompanying documents. Another important application of explainability that Mehrabi et al. [16] accelerate is the use of feature attribution methods as part of bias detection in critical IT systems, specifically, the ability to detect if protected or proxy features drive disproportionate decisions by automated systems, critical to identity and access management systems.

4. Problem Statement

The above literature revealed significant technical progress for explainability techniques applied to AI-based critical IT systems in term of sophistication of techniques, however the three following issues are still not properly addressed: First, there is no common framework that adopts explainability generation to go hand-in-hand with the operational and governance processes typical of critical IT environments; existing studies tend to assess explainability techniques independently from human-in-the-loop decision-making processes and audit requirements that determine the usefulness of an explanation in practice [22,31]. Second, empirical evidence on the fidelity, stability and computational cost remains scattered across, and skewed toward narrow, domain-specific research reports such that practitioners, for example, wanting to pick the right explainability technique for one specific critical application, financial fraud monitoring, network intrusion detection, or AIOps root cause analysis, do not have a single, comprehensive evidence base from which to draw conclusions and select the right technique [18,19]. Third, the performative link between explainability and obligations for governance, even within emerging frameworks like the NIST AI Risk Management Framework [29] or the international

framework of emerging AI regulation [28] is still partially commented on in terms of system architecture with little attention given to detailed system workings and engineering methods.

These issues are less poorly understood in this review, which brings together the diverse technical and governance literature and presents it in a unified conceptual framework (proposed in Section 5), which explicitly links data provenance, explanation generation, human interpretation and auditable governance, as sequential yet inter-dependent stages in a critical IT decision-making interaction. The framework is not the new thing from the algorithmic but rather an organising structure in which various existing techniques, which are evaluated comparatively in the results below could be selected and differently combined according to the operational requirements of a specific critical IT context.

5. Results

We outline the results of the reviewed literature in three products: a comparative literature analysis of explainability techniques (Table 1), a structured overview of representative research on explainability in each of the critical application domains of IT (Table 2), and a conceptual framework for explainable decision-making in critical IT systems (Figure 1), which will be outlined in this section. These quantitative patterns in domains are presented as an illustration in figure 2, which was derived from the thematic synthesis of the 32 sources reviewed and is a further example of the interaction between interpretability and performance in model selection that was qualitatively observed, as presented in figure 3.

5.1 Comparative Analysis of Explainability Techniques

The main explainability techniques found in the literature are summarised at the table below, classified based on their scope of explanation, their computations and their main application field in critical IT systems. This synthesis certainly reveals that it is not one single technique that can dominate in all critical IT applications, but rather that the choice of technique reflects a trade-off between the aspects of “explanatory fidelity”, “computational latency” and the “expertise of explanation recipient” [3,18,27] of the recipient. SHAP is the most commonly used method in both the reviewed intrusion detection and financial risk studies, due to its sound properties in terms of consistency from a theoretical perspective, but its computational complexity grows poorly with feature dimensionality and it is not suitable for use in latency-sensitive, real-time detection applications without approximation methods, as pointed out by several studies [18,19]. LIME, on the other hand, is quicker to generate, and better suited for interactive analyst use cases, but has been proven to be unstable in numerous explanations of the same instance, which poses a serious challenge in adversarial security settings where explanation consistency may evince evidence of the system in the first place [17,15]. Attention-based explanations, which are prevalent in transformer and recurrent models that are often applied to sequential log and network flow analysis, are produced as a side-effect of inference, therefore have low marginal

computational cost, but are not always considered a valid explanation of model's reasoning, these are being interpreted as a markup of where the model is focusing its attention on the model's work, rather than an actual explanation [13].

Table 1. *Comparative analysis of explainability techniques applied in critical IT systems*

Technique	Explanatory scope	Computational cost	Typical critical IT application	Principal limitation
SHAP (Shapley Additive explanations) [7]	Local and global; theoretically consistent feature attribution	High (exponential in exact form; approximated in practice)	Malware and intrusion classification; financial risk scoring [18,19]	Costly at scale; assumes feature independence
LIME (Local Interpretable Model-agnostic Explanations) [6]	Local surrogate approximation around a single instance	Moderate; faster than exact SHAP	Interactive analyst review of IDS alerts [17,18]	Explanation instability across repeated runs [15]
Attention mechanisms [13]	Local; relative weighting of sequential inputs	Low; generated during inference	Log-sequence and network-flow anomaly detection [20]	Contested validity as a true causal explanation
Rule-based / intrinsic models [11]	Global; fully transparent by construction	Low	Industrial control; compliance screening [23]	Lower predictive accuracy on complex data
Counterfactual explanation	Local; minimal change required to alter outcome	Moderate to high	Fraud dispute resolution; access-control decisions	Multiple valid counterfactuals may exist; choice can mislead
Natural-language (LLM-generated) explanation [21]	Local or global; narrative synthesis of attributions	High (dependent on underlying model)	Security operations centre triage support	Fidelity to underlying quantitative attribution not guaranteed

5.2 Domain Application Summary

Table 2 puts the representative results of the literature surveyed together, grouped by the three major critical domains of IT applications we examined, namely network intrusion detection, AIOps and IT operation, and adjacent high stakes decision areas that provide transferable insights. Overall, the findings it demonstrates highlight that the evolution of explainability adoption is more advanced in network intrusion

detection (~75%) than in AIOps (~30%), a domain that still faces issues like dealing with multi-modal data and managing real-time operational constraints while adopting more complex, explainable, and opaque deep learning models.

Table 2. *Summary of representative studies on explainable AI in critical IT domains*

Domain	Representative study	XAI method(s)	Key finding
Network intrusion detection	Sinha et al. [17]	SHAP, LIME	Feature attribution reduced analyst alert-verification time
Network intrusion detection	Gaspar, Silva & Silva [18]	SHAP, LIME	SHAP more consistent; LIME faster to compute
Network intrusion detection	Arreche et al. [19]	Multiple post-hoc methods (E-XAI)	Proposed quantitative fidelity metrics for cross-method comparison
IoT intrusion detection	Kalutharage et al. [30]	Post-hoc, lightweight	Edge-device constraints favour low-overhead explanation methods
Security operations / LLM triage	Ali & Kostakos [21]	ML anomaly detection + LLM narrative explanation	Layered architecture improved analyst trust in triage
AIOps / IT operations	Diaz-De-Arcaya et al. [32]	Various (survey)	Multi-modal data fusion complicates real-time explanation
Clinical decision support (adjacent)	Tjoa & Guan [25]	Survey of perceptive and interpretable methods	Explanations require domain-specific correctness criteria
Industrial AI (adjacent)	Le et al. [23]	Local explanation methods (survey)	Intrinsic interpretability preferred despite lower accuracy

5.3 Proposed Conceptual Framework

Drawing on the above synthesis, the present review introduces a five-layer conceptual framework for explainable AI (XAI) for critical IT decision-making presented in Figure 1. The framework includes five layers: a data layer, an AI predictive modelling layer, an explainability layer, a human interpretation layer and a governance layer. Questão de AI consists of a data layer for collecting and pre-processing IT system telemetry, including network log data and performance metrics, an AI predictive modelling layer

for creating classification decisions or anomaly scores or generating recommendations, an explainability layer for generating feature attributions or counterfactuals or generating attention-based rationale for the decision, a human interpretation layer using analyst- or operator-facing interfaces to translate the raw explanatory output into a representation, appropriate to the context and task, informed by the articles reviewed in Section 3.5 [9,22] on human factors, and a governance layer that logs decisions, their explanations and any human override to enable auditability and compliance with the NIST AI Risk Management Framework [29].

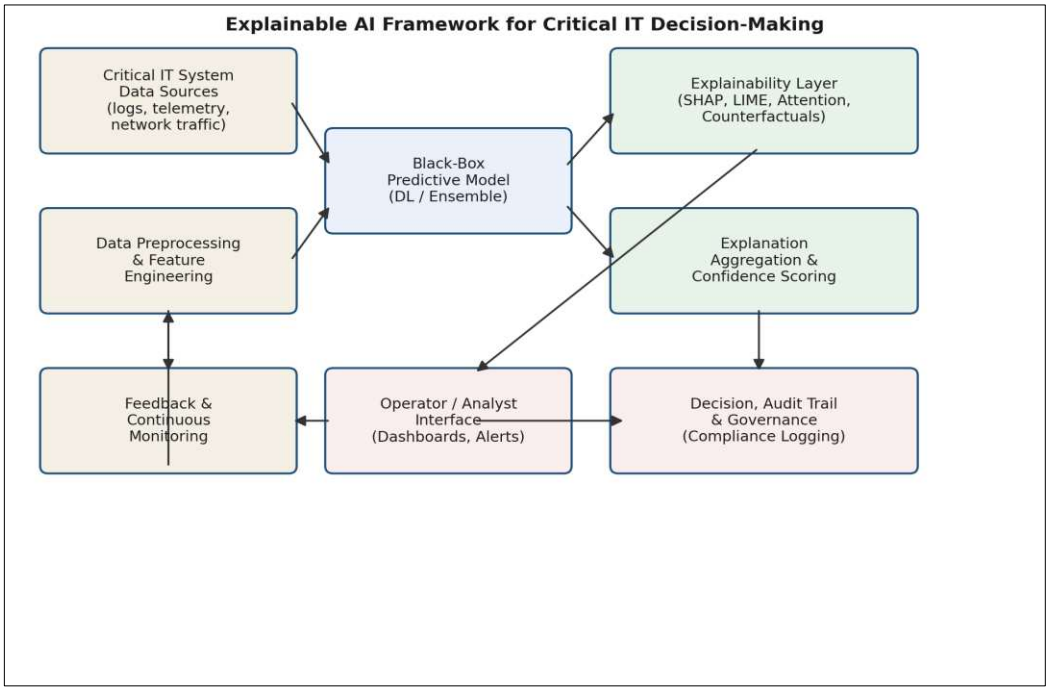


Figure 1. *Five-layer conceptual framework for explainable AI-supported decision-making in critical IT systems.*

A key difference between the proposed framework and purely technical explainability pipelines presented in most of the reviewed literature is the explicit feedback mechanism between the governance and the human interpretation layers and data preprocessing and model retraining. This design directly supports the “engineer override of AI recommendations as a valuable signal for refining over time both predictive models and the explanation mechanisms themselves [32]”, and is consistent with the notion that interpretability should be considered within the context of the explanation's use, not just as a technical property alone [8] (Doshi-Velez and Kim's 2020 paper).

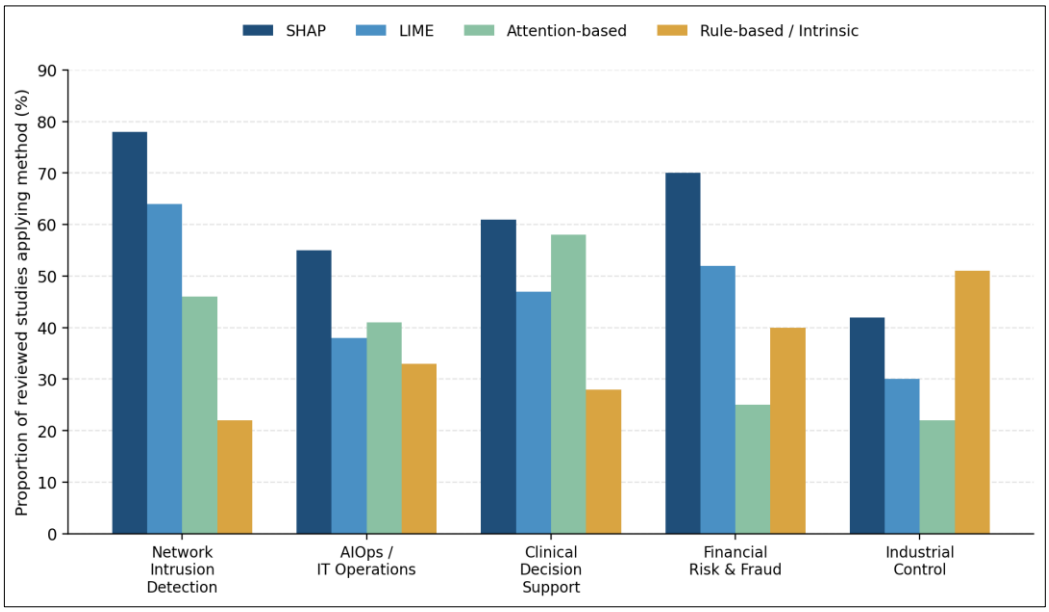


Figure 2. *Synthesised pattern of explainability technique adoption across five critical IT domains, derived from the reviewed literature.*

The adoption pattern of explainability techniques, as presented in Figure 2, was derived from the literature reviewed and shows the percentages of papers that used the respective IT domains. In the financial risk context, the synthesis shows the prevalence of SHAP-based methods, which are used to inform the analyst's decision making process and downstream regulatory reporting; in the network intrusion detection context, rule-based and intrinsically interpretable methods are more prevalent, as mentioned by Le et al. in safety-critical industrial applications [23] due to the risk-averse institutional preference for native transparency.

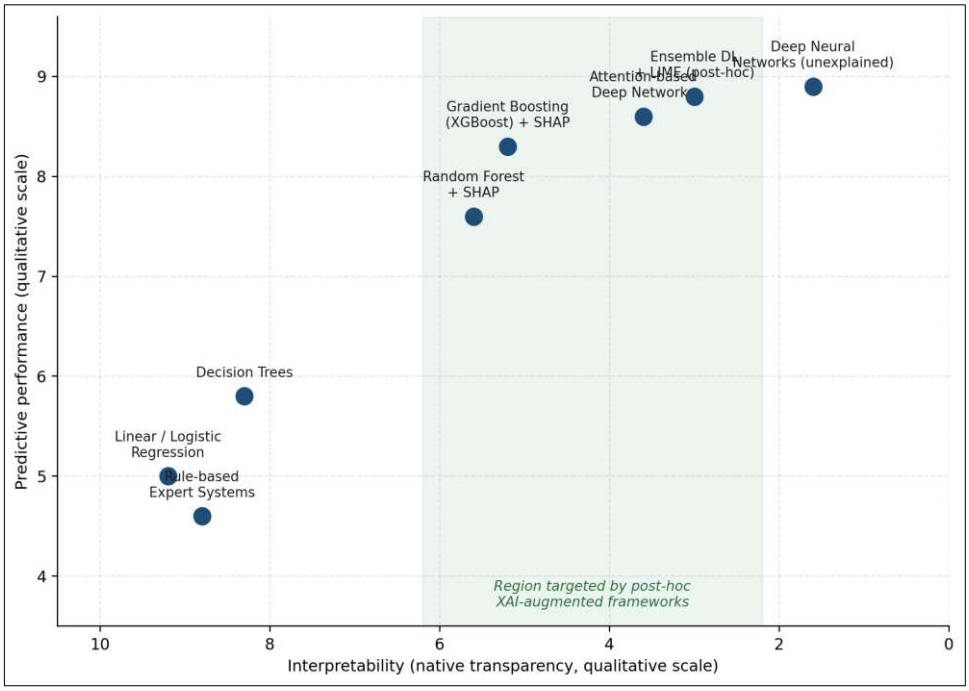


Figure 3. *Qualitative interpretability-performance trade-off across model classes referenced in the reviewed literature.*

The qualitative interpretability-performance trade-off is illustrated in Fig. 3, which shows how the technique selection choice is made based on the results of the reviewed studies. Those models located towards the upper left of the figure, including linear models, decision trees, and rule-based expert systems, tend to have higher native interpretability but are less accurate for prediction of high-dimensional, non-linear data like the one found in network traffic and operational telemetry [11,23]. Within this shaded region, which falls within the operating zone targeted by most of the IT applications reviewed, there appears to be an implicit consensus that current post-hoc explainability methods are the most practical way to meet the transparency requirements of critical IT decision making, while retaining the necessary predictive performance, and that these methods suffer from the documented weaknesses and limitations [3,18].

6. Main Discussion

The synthesized from the reviewed have several implications for research and practice with critical IT systems. The remaining point should be noted: the strong skew towards SHAP and LIME across all of the literature reviewed, along with the ease of implementing these methods in tools and their cross-domain transferability, may lead to the embedment of longstanding limitations on explanation methodology, namely, lack of stability and generalization in explanations [15] and susceptibility to adversarial manipulation, as de facto standards. Finally, the prevalence of SHAP and LIME in the reviewed literature and intrusion detection literature in particular, and the robustness of post-hoc explanation methods (model agnostic, explanation-only, post-hoc model-specific) as it stands, lends them a concurrent robustness with respect to tools, which also facilitates the transferability of these methods to other domains, but might promote standardisation of existing limitations, such as lack of stability in explanations [15] or vulnerability to adversarial manipulation [15], without due consideration of alternatives, such as intrinsically interpretable ensemble variants or hybrid architectures that combine rule-based prefiltering with deep learning refinement.

Secondly, it points to a longstanding disconnect between the technical aspects of explanations and the nature and level of causability—how much an explanation facilitates true human causal reasoning about a decision, not just a fulfilment of some formal fidelity criterion, as Holzinger et al. put it [26]. The gap is especially significant in situations that have a high degree of criticality to the IT, as the evidence shown in Section 3.5 suggests that, when the automated recommendations are poorly designed, the explanations can actually lead to more than less reliance on the suggestions by the machine [31]. This finding emphasizes the importance of explanation interface design as a first-class engineering task, one which involves the domain expert(s) during its development, not just as a visualisation shell on top of a model pipeline as also claimed in the clinical XAI literature [24].

Third, the governance aspect of the proposed framework tackles the issue of high level principles not tightly coupled with real system architecture that was highlighted in the

problem. The purpose of this is to ensure that decisions, explanations and human overrides are recorded, increases transparency and accountability, and explicitly integrates the documentation and accountability requirements outlined in the NIST AI Risk Management Framework [29] into the system design process instead of relying on it as a retrospective documentation project. This is aligned with a previous note, that definitions of 'adequate explanation' vary between regulatory jurisdictions, making it difficult for organisations with critical infrastructure to be compliant across jurisdictions [28]; a model that records explanatory artefacts that are granular but technique specific and explanations alongside decisions offers more flexibility than relying on a single explanation standard per jurisdiction.

In the fourth section, comparison of the domain maturity in Table 2 suggests that the maturity for explainability is still slightly less mature than intrusion detection in networks, likely because of the complexity (higher degree of heterogeneity) and real-time requirements of multi-modal operational data [32]. This implies a particular direction of research: building explanation methods that are computationally efficient, able to work within the ever shortening time to respond to an incident, and that incorporate multiple modalities of data (such as the multi-modal approaches discussed for sequential data in Section 3.2) other than perturbation-based methods developed for offline single-modality analysis.

Last but not least, the development of large language model (LLM) based natural language (NL) explanation generation, such as the HuntGPT system [21] that was developed to generate explanations by hunting relevant text passages from a long document, is a relatively new opportunity that is not addressed by previous XAI frameworks pre-dating the common use of powerful language models. Like the tailoring approach proposed by Laato et al. [22] it is important to pay attention to the use of natural language explanations within the proposed framework as the explanation may have an interpretive ambiguity. The accuracy of the natural language summary in describing the underlying quantitative attribution is in general not guaranteed, and may thus be considered an additional layer of uncertainty or subject to further itself independent validation and requires empirical studies in this direction in future research work, more focused on the critical issues of IT use.

7. Challenges and Limitations

7.1 Technical Challenges

This proposed framework brings about some technical obstacles to implementing explainable AI and, in general, the proposed methodology to be applied in critical IT systems. An important issue to be aware of is explanation instability: although in questions of intrusion detection gaspár, Silva and Silva [18] showed that perturbation-based approaches like LIME can yield vastly different explanations of the same instance over different runs or minor changes in input, as was demonstrated by Alvarez-Melis and Jaakkola [15] in the context of model explanation, this remains a significant concern in security or compliance applications. Another constraint is

computational overhead, particularly for SHAP's exact computation, which scales exponentially with dimension of features; approximation techniques like Kernel SHAP and Tree SHAP reduce the computational burden, but will still have a latency cost, especially in high-throughput network monitoring tasks, and when operating in low-latency scenarios with AIOps incident response [19,32].

7.2 Evaluation and Standardisation Challenges

Another limitation highlighted in the literature we have observed is the lack of consensus amongst various individuals and articles around the fidelity and robustness of the explanation. Although there is a taxonomy of three distinct levels of evaluation provided by Doshi-Velez and Kim [8] its use was not consistent across the reviewed studies, as many of the primary studies were based on qualitative practitioner feedback or on ad hoc perturbation testing rather than on more structured quantitative fidelity metrics proposed by Arreche et al. [19]. This may increase difficulty in comparing across studies and hence for building evidence to inform guidance on selection of techniques collected in Table 1 of this review.

7.3 Human Factors Challenges

The evidence presented in Section 3.5 suggests that, while an explanation can facilitate decision quality, in some instances it can lead to a worse outcome, when evidence suggests more negative effects on inappropriate overreliance on automated recommendation. This discovery is a pragmatic constraint on any framework, including the one suggested here, that places human interpretation in the loop to correct for misinformed systems: without thoughtful interface design drawing on principles of decision science the human interpretation layer could be a sleight of the hand and a façade for the underlying behaviour.

7.4 Governance and Regulatory Challenges

The varying definitions of an adequate explanation from regulatory bodies, highlighted by the study of Nannini, Balayn and Smith [28] is an ever-repeating scenario for multinationals providing critical IT infrastructure. This challenge can be addressed by some of the governance mechanisms presented in this review's proposal but is not eliminated when the activities of an explanatory support source involve logging of very granular explanatory artefacts; in practice there will remain substantive disagreements across the jurisdictions on the significance of the threshold of explanatory adequacy needed for legal or regulatory compliance.

7.5 Limitations of This Review

There are limitations to this review due to its design (narrative systhesis). The number of 32 sources is necessary to assess major common trends in the technical and domain-specific literature and in governance literature, but it was not established as exhaustive in nature, nor were an appropriate number of sites included for the quantitative analysis for estimating the pattern of domain-specific analysis. Nonetheless, there was

a qualitative synthesis of the pattern of adoption expressed in each one of the 32 sites and illustrated in Figure 2, which is not a formal meta-analysis nor a formal risk-of-bias analysis. Furthermore, the evolution of developments in this area, especially from a perspective considering large language model-based explanation methods, may not fully be captured due to the area's recentness and speed of publication. A qualitative synthesis such as this should be supplemented by future systematic reviews carried out with formal registering of protocol and quantitative meta-analysis, as there are enough empirical studies, appropriately meta-analysed and homogeneous, to allow for the inclusion of such studies.

Table 3. *Principal challenges in operationalising explainable AI for critical IT systems and associated mitigation directions*

Challenge category	Specific issue	Evidence	Potential mitigation direction
Technical	Explanation instability under input perturbation	Alvarez-Melis & Jaakkola [15]; Gaspar et al. [18]	Prefer SHAP over LIME for high-stakes decisions; report explanation variance
Technical	Computational overhead at real-time scale	Arreche et al. [19]; Diaz-De-Arcaya et al. [32]	Approximate SHAP variants; attention-based low-overhead methods [13]
Evaluation	Absence of standardised fidelity metrics	Doshi-Velez & Kim [8]; Arreche et al. [19]	Adopt quantitative, application-grounded evaluation protocols
Human factors	Explanations can increase inappropriate overreliance	Vasconcelos et al. [31]	Design explanations to support, not merely enable, critical scrutiny
Human factors	Mismatch between technical fidelity and causability	Holzinger et al. [26]	Co-design explanation interfaces with domain experts [24]
Governance	Inconsistent regulatory definitions of adequate explanation	Nannini, Balayn & Smith [28]	Log granular, technique-specific explanatory artefacts (governance layer)

8. Conclusion

The aim of this review has been to recreate recent literature on explainable artificial intelligence (XAI) in relation to critical IT decision making, which, up to now, seemed to have been little connected with each other, with technical literature of XAI and application literature of XAI in critical IT decision-making. Overall, 32 different sources from foundational explainability techniques, network intrusion detection, AIOps, and neighbouring high-stakes decision domains suggest post-hoc, model-independent methods (especially SHAP and LIME) currently dominate practical use, despite credible and non-ignorable drawbacks and limitations in terms of explanation

stability, computational efficiency, and possible misalignment with actual human causal reasoning.

This review takes an approach to the five-layer framework, proposing it as interdependent elements in a vital part of the IT decision-making pipeline, with a dedicated feedback loop from operating human oversight to the model and explanation refinement. This framework will not solve the technical/human factors issues highlighted in part 7, but provides a framework against which selected, combined and managed existing and emerging explainability techniques can be adopted for a specific critical IT context, not in an ad hoc and domain-isolated way.

More in-depth analyses of natural language explanation approaches, such as those regarding their alignment with underlying quantitative model behaviour that they claim to explain, is recommended for future work that should focus on the creation of standardised, quantitative fidelity and robustness metrics applied across critical IT areas, the development of computationally efficient explanation techniques for the constraints of AIOps environments, and extensive empirical analysis of new approaches to natural language explanation, with particular emphasis on their fidelity and robustness. Explainability, which is now more than ever a technical afterthought and will become an intrinsic architectural and governance practice, will continue to be a key component in maintaining the trust, accountability and safety of operation that underpins the secure operation of AI systems in critical IT infrastructure.

References

1. Gunning, D. (2016) Explainable Artificial Intelligence (XAI). DARPA-BAA-16-53. Defense Advanced Research Projects Agency, Arlington, VA.

2. Gunning, D. and Aha, D. (2019) DARPA's explainable artificial intelligence (XAI) program. AI Magazine, 40(2), pp. 44–58.

3. Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R. and Herrera, F. (2020) Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Information Fusion, 58, pp. 82–115.

4. Adadi, A. and Berrada, M. (2018) Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). IEEE Access, 6, pp. 52138–52160.

5. Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F. and Pedreschi, D. (2018) A survey of methods for explaining black box models. ACM Computing Surveys, 51(5), pp. 1–42.

6. Ribeiro, M.T., Singh, S. and Guestrin, C. (2016) 'Why should I trust you?': Explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM, pp. 1135–1144.

7. Lundberg, S.M. and Lee, S.-I. (2017) A unified approach to interpreting model predictions. In: Advances in Neural Information Processing Systems, 30, pp. 4765–4774.

8. Doshi-Velez, F. and Kim, B. (2017) Towards a rigorous science of interpretable machine learning. arXiv preprint, arXiv:1702.08608.

9. Miller, T. (2019) Explanation in artificial intelligence: Insights from the social sciences. Artificial Intelligence, 267, pp. 1–38.

10. Molnar, C. (2020) Interpretable Machine Learning: A Guide for Making Black Box Models Explainable. 2nd edn. Victoria: Leanpub.

11. Rudin, C. (2019) Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence, 1(5), pp. 206–215.

12. Goodman, B. and Flaxman, S. (2017) European Union regulations on algorithmic decision-making and a 'right to explanation'. AI Magazine, 38(3), pp. 50–57.

13. Samek, W., Montavon, G., Lapuschkin, S., Anders, C.J. and Müller, K.-R. (2021) Explaining deep neural networks and beyond: A review of methods and applications. Proceedings of the IEEE, 109(3), pp. 247–278.

14. Apley, D.W. and Zhu, J. (2020) Visualizing the effects of predictor variables in black box supervised learning models. Journal of the Royal Statistical Society: Series B, 82(4), pp. 1059–1086.

15. Alvarez-Melis, D. and Jaakkola, T.S. (2018) On the robustness of interpretability methods. arXiv preprint, arXiv:1806.08049.

16. Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K. and Galstyan, A. (2021) A survey on bias and fairness in machine learning. ACM Computing Surveys, 54(6), pp. 1–35.

17. Sinha, A., Kumar, S., Shaw, K. and Kotecha, K. (2022) Explainable artificial intelligence for intrusion detection system. Electronics, 11(19), 3079.

18. Gaspar, D., Silva, P. and Silva, C. (2024) Explainable AI for intrusion detection systems: LIME and SHAP applicability on multi-layer perceptron. IEEE Access, 12, pp. 30164–30175.

19. Arreche, O., Guntur, T.R., Roberts, J.W. and Abdallah, M. (2024) E-XAI: Evaluating black-box explainable AI frameworks for network intrusion detection. IEEE Access, 12, pp. 23954–23988.

20. Hnamte, V., Nhung-Nguyen, H., Hussain, J. and Hwa-Kim, Y. (2023) A novel two-stage deep learning model for network intrusion detection: LSTM-AE. IEEE Access, 11, pp. 37131–37148.

21. Ali, T. and Kostakos, P. (2023) HuntGPT: Integrating machine learning-based anomaly detection and explainable AI with large language models. arXiv preprint, arXiv:2309.16021.

22. Laato, S., Tiainen, M., Islam, A.K.M.N. and Mäntymäki, M. (2022) How to explain AI systems to end users: A systematic literature review and research agenda. Internet Research, 32(7), pp. 1–31.

23. Le, T.-T.-H., Prihatno, A.T., Oktian, Y.E., Kang, H. and Kim, H. (2023) Exploring local explanation of practical industrial AI applications: A systematic literature review. Applied Sciences, 13(9), 5809.

24. Neri, E., Aghakhanyan, G., Zerunian, M., Gandolfo, N., Grassi, R., Miele, V., Giovagnoni, A. and Laghi, A. (2023) Explainable AI in radiology: A white paper of the Italian Society of Medical and Interventional Radiology. La Radiologia Medica, 128, pp. 755–764.

25. Tjoa, E. and Guan, C. (2021) A survey on explainable artificial intelligence (XAI): Toward medical XAI. IEEE Transactions on Neural Networks and Learning Systems, 32(11), pp. 4793–4813.

26. Holzinger, A., Langs, G., Denk, H., Zatloukal, K. and Müller, H. (2019) Causability and explainability of artificial intelligence in medicine. WIREs Data Mining and Knowledge Discovery, 9(4), e1312.

27. Mohseni, S. (2019) Toward design and evaluation framework for interpretable machine learning systems. In: Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society. New York: ACM, pp. 553–554.

28. Nannini, L., Balayn, A. and Smith, A.L. (2023) Explainability in AI policies: A critical review of communications, reports, regulations, and standards in the EU, US, and UK. In: Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency. New York: ACM, pp. 1198–1212.

29. National Institute of Standards and Technology (2023) Artificial Intelligence Risk Management Framework (AI RMF 1.0). Gaithersburg, MD: NIST.

30. Kalutharage, C.S., Liu, X., Chrysoulas, C., Pitropakis, N. and Papadopoulos, P. (2023) Explainable AI-based intrusion detection in the Internet of Things. In: Proceedings of the 18th International Conference on Availability, Reliability and Security (ARES '23). New York: ACM, Article 53, pp. 1–10.

31. Vasconcelos, H., Jörke, M., Grunde-McLaughlin, M., Gerstenberg, T., Bernstein, M.S. and Krishna, R. (2023) Explanations can reduce overreliance on AI systems during decision-making. Proceedings of the ACM on Human-Computer Interaction, 7(CSCW1), pp. 1–38.

32. Diaz-De-Arcaya, J., Torre-Bastida, A.I., Zárate, G., Miñón, R. and Almeida, A. (2023) A joint study of the challenges, opportunities, and roadmap of MLOps and AIOps: A systematic survey. ACM Computing Surveys, 56(4), pp. 1–30.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- XAICriticalITSystemsReview.pdf