Adaptive Multi-Granularity Explainable AI for Real-Time Cybersecurity Threat Detection: Bridging the Trust-Latency-Novelty Gap: A Conceptual Framework

Chandras Batheja [1], Tathagat Suryavanshi [2], Tanmay Suryavanshi [3]
[1] *Corresponding Author, Research Scholar (Ph.D. Student), Indian Management School and Research Centre, Mumbai-400705, India*
[2]*Bachelor of Technology in Computer Science, Galgotias College of Engineering and Technology (Dr. APJ Abdul Kalam Technical University)*
[3] *Bachelor of Technology in Computer and Communication Engineering, Manipal University, Jaipur*

***Abstract*—Machine learning models deployed for intrusion detection, malware classification, and phishing identification have achieved high predictive accuracy, yet their opaque decision-making limits adoption in security operations centers (SOCs), where analysts must justify actions to auditors, regulators, and incident-response teams within seconds. Explainable Artificial Intelligence (XAI) has been proposed as a remedy, but the cybersecurity-XAI literature remains dominated by static, post-hoc explanation techniques applied to offline benchmark datasets, evaluated primarily for fidelity to the underlying model rather than for usefulness to the humans who must act on them. This paper identifies four interlocking gaps in the current literature: the absence of stakeholder-adaptive explanation granularity, the neglect of explanation robustness under adversarial perturbation, the mismatch between real-time detection latency budgets and the computational cost of explanation generation, and the unresolved question of how to explain predictions for novel or zero-day attacks that fall outside a model's learned taxonomy. To address these gaps, we propose Adaptive Multi-Granularity Explainable AI (AMG-XAI), a conceptual framework that couples a novelty-aware detection layer with a stakeholder-conditioned explanation generator and an adversarial-robustness auditor, operating within an explicit latency budget. We detail the framework's architecture, propose an evaluation methodology spanning fidelity, robustness, latency, and human-centered usability metrics, and discuss the implications for SOC workflows, regulatory compliance, and future research. The contribution is deliberately framed as a conceptual and methodological one: it synthesizes a fragmented literature into a coherent research agenda and offers a concrete, testable framework rather than a single-dataset empirical result.**

***Index Terms*—Explainable AI, cybersecurity, intrusion detection, SHAP, adversarial robustness, security operations centers, human-computer trust, zero-day detection.**

I. INTRODUCTION

The last decade has seen machine learning (ML) and deep learning (DL) become the dominant paradigm for detecting cyberattacks, from network intrusion detection systems (NIDS) to malware classifiers, phishing filters, and behavioral anomaly detectors for insider threats. These models routinely outperform signature-based and rule-based systems on benchmark datasets, particularly against previously unseen attack variants. Yet their adoption inside real security operations centers has been slower and more contested than the accuracy figures would suggest, and a recurring reason cited by practitioners is opacity: a model that flags a connection as malicious without indicating why offers little that a SOC analyst can act on, defend to an auditor, or use to update detection rules.

Explainable AI (XAI) has emerged as the proposed bridge between predictive performance and operational trust. Techniques such as SHapley Additive exPlanations (SHAP), Local Interpretable Model-agnostic Explanations (LIME), attention visualization, and counterfactual explanation have all been transplanted from computer vision and general tabular-data settings into cybersecurity applications,

typically to explain why a particular flow, packet, or file was classified as malicious. This transplantation has produced a fast-growing body of work, and several surveys now catalogue dozens of XAI-for-cybersecurity papers published since 2019.

This growth in publication volume, however, has not been matched by a corresponding maturation of the underlying research questions. Most studies follow a similar template: take an established XAI technique, apply it to a known intrusion-detection or malware dataset, and report that the resulting feature attributions are plausible to the authors. Comparatively little attention has been paid to whether these explanations remain stable when an adversary deliberately manipulates the input, whether they can be generated within the latency budgets that real-time detection pipelines require, whether different consumers of an explanation (a Tier-1 analyst triaging alerts, a Tier-3 threat hunter reverse-engineering a campaign, a compliance officer preparing a regulatory report) need fundamentally different representations of the same underlying evidence, or how a model should explain a prediction about an attack type it has never been trained to recognize.

This paper makes three contributions. First, it synthesizes the XAI-for-cybersecurity literature into a structured taxonomy organized by application domain and explanation technique, making explicit the assumptions each line of work relies on. Second, it identifies and argues for four specific, underexplored research gaps that emerge from this synthesis, each grounded in a concrete operational failure mode rather than a purely academic distinction. Third, it proposes Adaptive Multi-Granularity Explainable AI (AMG-XAI), a conceptual framework designed to address these gaps jointly rather than in isolation, together with an evaluation methodology that operationalizes fidelity, robustness, latency, and human usability as measurable, reportable criteria. The remainder of the paper proceeds as follows: Section 2 provides background on XAI techniques and their cybersecurity-specific constraints; Section 3 reviews related work; Section 4 develops the research gap in detail; Section 5 presents the AMG-XAI framework; Section 6 proposes an evaluation methodology; Section 7 discusses implications; Section 8 acknowledges limitations; and Section 9 concludes with directions for future work.

II. BACKGROUND: XAI TECHNIQUES AND CYBERSECURITY-SPECIFIC CONSTRAINTS

2.1 Families of Explanation Techniques

XAI methods are conventionally divided into intrinsically interpretable models and post-hoc explanation methods applied to black-box models. Intrinsically interpretable models, such as decision trees, rule lists, and generalized additive models, expose their reasoning directly but typically sacrifice some predictive accuracy relative to ensemble or deep learning approaches, which matters in a domain where missed detections carry high cost. Post-hoc methods instead approximate or probe a trained black-box model after the fact. Perturbation-based methods such as LIME fit a local surrogate model around a single prediction; game-theoretic methods such as SHAP attribute a prediction to input features using Shapley values from cooperative game theory; gradient-based methods such as saliency maps and Grad-CAM trace a prediction back to influential input regions in differentiable models; attention mechanisms expose learned weightings in sequence and graph models; and counterfactual methods identify the minimal change to an input that would flip the model's decision.

Each family carries distinct assumptions. Perturbation-based and game-theoretic methods are computationally expensive, often requiring hundreds to thousands of model evaluations per explanation, which is tolerable for offline forensic analysis but frequently incompatible with line-rate network traffic inspection. Gradient-based and attention-based methods are cheap to compute but only apply to differentiable architectures and can be misleading when attention weights do not correspond to causal influence on the output, a finding well documented outside the security domain. Counterfactual methods are intuitive to non-technical audiences but can be difficult to render meaningfully for high-dimensional, categorical network-flow features.

2.2 What Makes Cybersecurity Different

Applying XAI to cybersecurity is not a simple matter of porting general-purpose techniques onto a new dataset. Four properties of the domain distinguish it from the image-classification and credit-scoring settings in which most XAI methods were originally developed and validated. First, the input space is

adversarial by construction: unlike a loan applicant, an attacker actively seeks to manipulate the features a detector relies on, which means both the underlying model and any explanation derived from it are targets, not neutral objects of study. Second, detection is frequently a real-time, high-throughput task, where a NIDS may need to classify tens of thousands of flows per second, leaving a strict computational budget for any explanation that is generated inline rather than after the fact. Third, the consumers of an explanation are heterogeneous and hierarchical, ranging from automated response systems, to junior analysts under time pressure, to senior threat hunters conducting deep forensic work, to auditors who require documentation independent of technical detail. Fourth, and perhaps most consequentially, a meaningful fraction of security-relevant events are novel by definition: zero-day exploits and previously unseen malware families are exactly the cases a defender most needs explained, yet they are, by construction, underrepresented or absent in the training distribution that any explanation method implicitly relies on to define what a 'normal' contributing feature looks like.

2.3 Regulatory and Organizational Context

The push toward explainability in cybersecurity is not purely technical; it is increasingly shaped by regulatory and organizational pressure. Data-protection and AI-governance frameworks in multiple jurisdictions now expect that automated decisions with material consequences for individuals or organizations be accompanied by some form of rationale, and sector-specific guidance for critical infrastructure and financial services has begun to reference explainability explicitly as an expectation for automated security controls. Within organizations, SOC leadership must additionally justify the cost and false-positive burden of ML-based detectors to executives who are not machine learning specialists, which creates a second, internal audience for explanation distinct from the analyst who reads it during triage. These pressures mean that any practical XAI framework for cybersecurity must be evaluated not only on technical fidelity but on its ability to satisfy audiences with very different technical backgrounds and very different questions in mind, reinforcing the motivation for stakeholder-adaptive design developed later in this paper.

III. RELATED WORK

3.1 XAI for Network Intrusion Detection

A substantial share of the literature applies SHAP or LIME to gradient-boosted trees or deep neural networks trained on intrusion-detection benchmarks such as NSL-KDD, CICIDS2017, and UNSW-NB15, reporting feature-attribution rankings that highlight flow duration, packet-size statistics, and protocol flags as recurring top contributors to malicious classifications. These studies generally validate explanations qualitatively, by checking that the highlighted features are consistent with domain knowledge about the attack types present in the dataset, rather than through a quantitative, human-subject evaluation of whether the explanations improved an analyst's decision quality or speed. A smaller set of papers has explored attention-based architectures for sequence-of-packets or sequence-of-flows models, using the learned attention weights as a built-in explanation mechanism, and rule-extraction approaches that distill a trained deep model into a smaller decision-tree or rule-list surrogate for interpretability.

3.2 XAI for Malware and Phishing Detection

In malware classification, explanation work has concentrated on identifying which opcode sequences, API call patterns, or static-analysis features drive a classifier's decision, often visualized as heat maps over disassembled code or portable-executable header fields. Phishing and malicious-URL detection has similarly used LIME and SHAP to highlight which lexical or host-based features (URL length, presence of an IP address in the host field, domain age) contributed to a classification, with several studies proposing that such explanations could be surfaced directly to end users rather than only to analysts, as a way of building general public trust in automated filtering.

A related strand of work has examined graph-based representations of malware behavior, such as API call graphs or dynamic-analysis execution traces, and applied graph-specific explanation techniques to highlight influential nodes or subgraphs. These approaches are attractive because they align with how a human analyst already conceptualizes malware behavior, as a sequence of related actions rather than an unordered feature vector, but they remain

computationally heavier than flat-feature attribution methods and have rarely been evaluated for the latency constraints discussed in Section 2.2. Similarly, insider-threat and user-behavior-analytics research has begun to apply explanation techniques to sequence models trained on login and access-log data, though this sub-literature is smaller and even less standardized in its evaluation practices than network-intrusion or malware work.

3.3 General-Purpose XAI Foundations Relevant to This Work

The methodological backbone of most cybersecurity-XAI work traces to a small number of general-purpose techniques: Shapley-value attribution, local surrogate modeling, and counterfactual explanation, alongside foundational critiques of interpretability as a concept, which distinguish between explanations that are faithful to a model's internal computation and explanations that are merely persuasive to a human observer. A parallel and largely separate literature on adversarial machine learning has shown that both the predictions and the explanations of a model can be manipulated by an adversary who has access to the model or its outputs, including work demonstrating that two functionally near-identical models can be constructed to produce the same predictions while yielding arbitrarily different SHAP or LIME explanations, and that explanation methods themselves can be targeted to hide a model's reliance on a protected or sensitive feature.

3.4 Surveys and Systematizations

Several survey papers have organized the XAI-for-cybersecurity literature by explanation technique, by application domain, or by the type of underlying ML model, and most conclude with a call for standardized evaluation practices, since the field currently lacks a shared benchmark or shared set of metrics analogous to accuracy and F1-score for the predictive task itself. These surveys consistently note the shortage of human-subject studies, the near-total absence of adversarial evaluation of explanations, and the limited treatment of real-time computational constraints, but stop short of proposing a unifying framework that treats these as jointly, rather than separately, addressable problems. It is this gap between diagnosis and framework that the present paper aims to close.

3.5 Lessons from Adjacent High-Stakes Domains

Fields such as clinical decision support and credit-risk assessment have grappled with structurally similar tensions between model performance and explanation usability, and offer instructive parallels even though neither domain features an adaptive adversary in the way cybersecurity does. In clinical decision support, research has shown that clinicians of differing seniority and specialty prefer materially different explanation formats for the same underlying model output, which has motivated role-conditioned explanation interfaces broadly analogous to the stakeholder-adaptive renderer proposed in Section 5.3. In credit-risk assessment, regulatory requirements for adverse-action notices have pushed the field toward counterfactual and rule-based explanations that are actionable for the affected individual, a lesson that has not yet been systematically transferred to cybersecurity contexts where an equivalent actionable explanation might, for instance, tell a system owner precisely which configuration change would have prevented a flagged event from occurring. Both fields have also converged on the view that fidelity-only evaluation is insufficient and that human-subject studies measuring decision outcomes, not just explanation plausibility, are necessary to establish real-world value, a conclusion this paper extends to the cybersecurity setting in Section 6.5. What neither domain has had to confront, and what makes cybersecurity distinct even relative to these adjacent fields, is an adversary who can observe and adapt to the explanation mechanism itself, which is why Gap 2 has no direct analogue in the clinical or credit-risk XAI literature and requires cybersecurity-specific treatment.

IV. RESEARCH GAP

Synthesizing the literature reviewed in Section 3 against the domain-specific constraints identified in Section 2 surfaces four gaps that recur across application areas but have not, to date, been addressed within a single, integrated research agenda.

4.1 Gap 1: Absence of Stakeholder-Adaptive Explanation Granularity

Existing work overwhelmingly produces a single explanation format, most often a ranked list or bar chart of feature attributions, regardless of who will

consume it. A Tier-1 analyst triaging hundreds of alerts per shift needs a one-line, natural-language justification; a threat hunter needs the full attribution vector alongside the raw flow to support pivoting into related events; an auditor needs an explanation traceable to a documented policy or control. No framework in the reviewed literature conditions the depth, format, or vocabulary of an explanation on the identity or role of its consumer, despite this being a well-established requirement in the broader human-computer interaction literature on explanation.

4.2 Gap 2: Unexamined Robustness of Explanations to Adversarial Perturbation

Because attackers actively adapt to defensive measures, an explanation mechanism deployed in production becomes part of the attack surface: if an adversary can perturb an input to preserve a malicious classification's label while altering the resulting explanation, or conversely alter the explanation while preserving a benign classification, the explanation ceases to be trustworthy evidence. The adversarial-ML literature has demonstrated such attacks against explanation methods in general-purpose settings, but this threat has been almost entirely unaddressed within the cybersecurity-XAI literature specifically, where the adversary is not a hypothetical but the defining feature of the domain.

4.3 Gap 3: Mismatch Between Explanation Latency and Real-Time Detection Budgets

Perturbation-based and game-theoretic explanation methods, the most widely used in the surveyed literature, are computationally expensive relative to the millisecond-scale budgets available in line-rate network monitoring. Nearly all reviewed studies evaluate explanation quality offline, on batches of already-collected data, and do not report the latency cost of explanation generation, let alone whether that cost is compatible with the detection pipeline it purports to support. This leaves practitioners without guidance on which explanation techniques, if any, are viable for inline rather than retrospective use.

4.4 Gap 4: Unresolved Explanation of Novel and Zero-Day Threats

Attribution-based explanation methods implicitly explain a prediction in terms of the features and patterns the model learned during training. For a genuinely novel attack, whose behavior does not resemble any training-time attack class, the resulting explanation may cite features that are technically responsible for the model's output but are misleading as a causal account of the underlying threat, since the model itself has no learned representation of the true attack mechanism. The literature has not systematically distinguished between explaining a familiar attack and explaining an anomalous, out-of-distribution event, despite the latter being the higher-stakes case in practice.

These four gaps are not independent. A stakeholder-adaptive system (Gap 1) must decide, for a novel-attack case (Gap 4), whether to present a lower-confidence, exploratory explanation rather than a confident attribution; an adversarially robust explanation mechanism (Gap 2) is only useful if it can still be computed within the detection pipeline's latency budget (Gap 3). Addressing any one gap in isolation therefore risks solutions that do not compose. This observation motivates the integrated framework proposed in Section 5.

V. PROPOSED FRAMEWORK: ADAPTIVE MULTI-GRANULARITY EXPLAINABLE AI (AMG-XAI)

AMG-XAI is proposed as a layered architecture that sits alongside an existing ML-based detection pipeline rather than replacing it, so that it can, in principle, be retrofitted onto detectors already deployed in a SOC. The framework comprises five components, described below and summarized in Table 1.

5.1 Novelty-Aware Triage Layer

Before any explanation is generated, an incoming flagged event is first scored by a novelty detector, implemented as an out-of-distribution estimator (for example, a density-based or reconstruction-based model trained on the same distribution as the primary classifier) that outputs a scalar novelty score alongside the primary model's classification. Events with low novelty scores are routed to the standard attribution pipeline described below; events with high novelty scores are routed to an exploratory-explanation mode, in which the system explicitly flags that the explanation reflects statistical anomaly relative to training data rather than a confident causal account

tied to a known attack pattern, directly addressing Gap 4.

5.2 Latency-Budgeted Explanation Selector

A scheduling component tracks the real-time latency budget available for the current traffic load and selects among a tiered set of explanation methods: a cheap, always-available method (for example, feature-importance from the model's native structure, or a lightweight gradient-based method) for high-throughput periods, and a more expensive, higher-fidelity method (for example, sampled or approximate Shapley attribution) when spare computational budget exists or when an event is escalated for human review. This selector makes the latency-fidelity trade-off explicit and measurable rather than implicit and unreported, addressing Gap 3.

5.3 Stakeholder-Conditioned Explanation Renderer

The same underlying attribution output is rendered into role-specific formats via a template layer conditioned on a consumer profile: an analyst profile receives a short natural-language summary referencing the top two or three contributing factors; a threat-hunter profile receives the full attribution vector, the raw flow or artifact, and links to related historical events; an audit profile receives a structured record mapping the explanation to the specific detection policy or control it satisfies, suitable for compliance documentation. This directly operationalizes Gap 1.

5.4 Adversarial-Robustness Auditor

Before an explanation is surfaced, it is passed through a lightweight consistency check that perturbs the input within a bounded, semantically valid neighborhood (for example, feature perturbations consistent with an attacker's realistic degrees of freedom, such as padding a payload or altering timing, rather than arbitrary perturbation of every feature) and verifies that the resulting explanation does not change its top contributing features beyond a defined stability threshold. Explanations that fail this check are flagged as low-confidence and are not used to auto-generate detection rules, directly addressing Gap 2.

5.5 Feedback and Rule-Refinement Loop

Analyst feedback on the usefulness and accuracy of a rendered explanation is logged and used to periodically retrain the novelty detector's threshold and the explanation selector's tiering policy, closing the loop between operational use and framework calibration. This component is intended to keep the system's behavior aligned with evolving analyst needs and evolving attacker behavior over time, rather than treating explanation quality as a static, one-time property established at deployment.

Table 1 summarizes how each AMG-XAI component maps onto the corresponding research gap from Section 4.

Table 1. Mapping of AMG-XAI components to identified research gaps.

AMG-XAI component	Research Gap
Novelty-Aware Triage Layer	Gap 4 (explaining novel/zero-day threats)
Latency-Budgeted Explanation Selector	Gap 3 (real-time latency constraints)
Stakeholder-Conditioned Explanation Renderer	Gap 1 (stakeholder-adaptive granularity)
Adversarial-Robustness Auditor	Gap 2 (explanation robustness)
Feedback and Rule-Refinement Loop	cross-cutting calibration across all four gaps

5.6 Illustrative Walkthrough

A concrete, hypothetical scenario helps clarify how the five components interact. Consider a network flow flagged by the primary classifier as consistent with command-and-control beaconing. The novelty-aware triage layer first checks the flow against the training distribution; if its novelty score is low, indicating the flow resembles previously seen beaconing patterns, it proceeds to standard attribution. The latency-budgeted selector checks current traffic load: during a period of high throughput, it computes a cheap, native feature-importance explanation citing periodic timing intervals and an unusual destination-port pattern; during a quieter period, or once the event is escalated by a Tier-1 analyst, the more expensive sampled-Shapley explanation is computed instead, adding finer-grained attribution across a wider feature set. Before either explanation is surfaced, the robustness auditor perturbs the flow's timing and payload-size features within realistic attacker degrees of freedom

and confirms that the same top features remain dominant; had they not, the explanation would be flagged as unstable and withheld from automatic rule generation, though still shown to the analyst with a caveat. Finally, the renderer produces two outputs from the same underlying attribution: a one-line summary for the Tier-1 analyst's queue, noting that the flow was flagged for periodic beaconing to an uncommon port similar to prior confirmed incidents, and a full attribution record, linked to related historical flows, for the threat hunter who later investigates the same host. If the flow's novelty score had instead been high, the system would skip confident attribution altogether and instead present the specific statistical properties that made the flow anomalous relative to the training distribution, explicitly labeled as exploratory rather than confirmed, giving the threat hunter a starting point for manual investigation rather than a potentially misleading causal claim.

VI. PROPOSED EVALUATION METHODOLOGY

Because this paper's contribution is framework-level, we propose, rather than report, an evaluation methodology intended to make AMG-XAI and comparable systems empirically testable and to establish shared metrics the field currently lacks.

6.1 Datasets

A robust evaluation should combine established intrusion-detection benchmarks (CICIDS2017, UNSW-NB15, NSL-KDD) for comparability with prior work, a malware or phishing dataset to test generalization across application domains, and a held-out set of attack classes deliberately excluded from training to simulate zero-day conditions for testing the novelty-aware triage layer.

6.2 Fidelity Metrics

Explanation fidelity should be measured using established proxies such as deletion and insertion curves (measuring how a model's confidence changes as top-attributed features are progressively removed or added) and comparison against the underlying model's own feature-importance structure where available, to verify that rendered explanations remain faithful to the model rather than merely plausible to a human reader.

6.3 Robustness Metrics

Robustness should be quantified as the rate at which top-k attributed features change under bounded, realistic adversarial perturbation, alongside the rate at which the Adversarial-Robustness Auditor correctly flags unstable explanations, evaluated against known explanation-targeted attack techniques adapted from the general adversarial-ML literature.

6.4 Latency Metrics

Latency should be reported per explanation tier (cheap versus expensive) under realistic traffic-load simulation, alongside the proportion of events that can be served the higher-fidelity explanation within a defined service-level budget (for example, under 50 milliseconds for inline use, with a longer budget permitted for escalated forensic review).

6.5 Human-Centered Usability Metrics

Finally, and most critically given the near-absence of human-subject evaluation in the reviewed literature, usability should be assessed through controlled studies with SOC analysts of varying seniority, measuring time-to-decision, decision accuracy, and self-reported trust when using stakeholder-conditioned explanations versus a single generic explanation format, ideally using a within-subjects design to control for individual analyst variation.

VII. DISCUSSION AND IMPLICATIONS

If validated empirically, a framework of this kind carries implications beyond individual SOC deployments. For regulators and standards bodies increasingly requiring documented rationale for automated security decisions, a stakeholder-conditioned audit output offers a concrete mechanism for compliance that does not require exposing raw model internals to non-technical reviewers. For SOC operations, explicit latency-tiered explanation generation offers a template for retrofitting explainability onto existing high-throughput detection pipelines without redesigning them from scratch. For the research community, treating novelty detection, robustness auditing, latency budgeting, and stakeholder adaptation as components of a single evaluated system, rather than as separate lines of inquiry, offers a template for closing the gap the

surveyed literature has repeatedly diagnosed but not resolved.

The framework also surfaces a subtler point about the nature of explanation in adversarial domains: an explanation's value is not fixed but contextual, depending on who is reading it, how much computational budget was available to produce it, and whether the underlying event resembles anything the model has previously seen. Treating explanation quality as a single scalar to be maximized, as much of the current literature implicitly does through fidelity metrics alone, obscures this contextual dependence and may explain why high-fidelity explanations in benchmark studies have not translated into demonstrated operational trust in deployed systems.

There is also an economic dimension worth noting. Computing higher-fidelity explanations for every flagged event at line rate is not free, and organizations with constrained security budgets will need guidance on where the marginal analyst-hours saved by better explanations justify the additional compute expenditure. A latency-tiered approach such as the one proposed here offers one way of making that trade-off explicit and adjustable rather than fixed by whichever explanation technique a development team happened to select, and future empirical work should attempt to quantify this trade-off directly, for instance by estimating the reduction in mean time-to-triage attributable to a given tier of explanation against its computational cost per event.

VIII. LIMITATIONS

This paper presents a conceptual framework and evaluation methodology rather than an empirical validation of AMG-XAI against the proposed metrics; the framework's practical performance, and the extent to which its components can be implemented without introducing new latency or robustness weaknesses of their own, remains to be established through the empirical program outlined in Section 6. The proposed novelty-detection and robustness-auditing components also introduce their own computational overhead and false-positive behavior, which must be characterized empirically rather than assumed away, and the stakeholder profiles described in Section 5.3 are illustrative rather than derived from a formal requirements study of any specific organization's analyst roles.

IX. CONCLUSION AND FUTURE WORK

Explainable AI has been widely proposed as the mechanism by which opaque, high-performing cybersecurity models can earn operational trust, but the literature applying XAI to cybersecurity has largely reproduced techniques developed for other domains without adapting them to the properties that make cybersecurity distinct: an adversarial input space, strict real-time constraints, a hierarchy of heterogeneous stakeholders, and a persistent need to explain genuinely novel events. This paper has argued that these four properties correspond to four specific, underexplored research gaps, and has proposed Adaptive Multi-Granularity Explainable AI as an integrated framework addressing them jointly, together with an evaluation methodology intended to give the field a shared basis for comparing future work. Future research should prioritize empirical implementation and human-subject evaluation of the proposed framework across multiple SOC contexts, extension of the adversarial-robustness auditor to cover a wider range of realistic attacker perturbation models, and investigation of whether stakeholder-conditioned explanation generation can be learned end-to-end rather than implemented through hand-authored templates, as proposed here for tractability. Only through this kind of integrated, operationally grounded evaluation can the field move from demonstrating that explanations can be generated to demonstrating that they are trusted, used, and effective.

REFERENCES

[1] Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., ... & Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Information Fusion, 58, 82–115.

[2] Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. arXiv preprint.

[3] Guo, W. (2023). Explainable artificial intelligence for cybersecurity: A literature survey. Annals of Telecommunications, 78, 1–20.

[4] Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions.

Advances in Neural Information Processing Systems, 30.

[5] Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?": Explaining the predictions of any classifier. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 1135–1144.

[6] Sharafaldin, I., Lashkari, A. H., & Ghorbani, A. A. (2018). Toward generating a new intrusion detection dataset and intrusion traffic characterization. Proceedings of the 4th International Conference on Information Systems Security and Privacy, 108–116.

[7] Slack, D., Hilgard, S., Jia, E., Singh, S., & Lakkaraju, H. (2020). Fooling LIME and SHAP: Adversarial attacks on post hoc explanation methods. Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, 180–186.

[8] Wachter, S., Mittelstadt, B., & Russell, C. (2017). Counterfactual explanations without opening the black box: Automated decisions and the GDPR. Harvard Journal of Law & Technology, 31, 841.

[9] Warnecke, A., Arp, D., Wressnegger, C., & Rieck, K. (2020). Evaluating explanation methods for deep learning in security. Proceedings of the IEEE European Symposium on Security and Privacy (EuroS&P), 158–174.

[10] Zhang, Z., Al Hamadi, H., Damiani, E., Yeun, C. Y., & Taher, F. (2022). Explainable artificial intelligence applications in cyber security: State-of-the-art in research. IEEE Access, 10, 93104–93139.