Position: the explainability paradox: challenges for XAI in malware detection and analysis

Li, R.; Gadyatskaya, O.

Position: The Explainability Paradox – Challenges for XAI in Malware Detection and Analysis

Rui Li
LIACS, Leiden University
Leiden, The Netherlands
r.li@liacs.leidenuniv.nl

Olga Gadyatskaya
LIACS, Leiden University
Leiden, The Netherlands
o.gadyatskaya@liacs.leidenuniv.nl

2024 IEEE European Symposium on Security and Privacy Workshops (EuroS&amp;PW) | 979-8-3503-6729-4/24/$31.00 ©2024 IEEE | DOI: 10.1109/EUROSPW61312.2024.00067

Abstract—**Malware poses a significant threat to global cybersecurity, with machine learning emerging as the primary method for its detection and analysis. However, the opaque nature of machine learning's decision-making process often leads to confusion among stakeholders, undermining their confidence in the detection outcomes. To enhance the trustworthiness of malware detection, Explainable Artificial Intelligence (XAI) is employed to offer transparent and comprehensible explanations of the detection mechanisms, which enable stakeholders to gain a deeper understanding of detection mechanisms and assist in developing defensive strategies. Despite the recent XAI advancements, several challenges remain unaddressed. In this paper, we explore the specific obstacles encountered in applying XAI to malware detection and analysis, aiming to provide a road map for future research in this critical domain.**

Index Terms—**Malware detection, explainable AI, XAI, malware explanations, XAI performance assessment.**

1. Introduction

Malware is one of the most important cybersecurity threats. It has a high impact on individuals, organizations, and national infrastructure. In 2023, more than 100 million samples of malware and potentially unwanted applications (PUA) were identified by AV-Test [3]. Machine learning (ML) is the key to effectively detecting these large numbers of malicious applications. Yet, these ML-based systems are usually not transparent and security analysts may find it difficult to understand the reasons why a certain sample is flagged as malicious. To improve the trustworthiness of malware detection systems and support the analysts, *Explainable AI* (*XAI*) plays a crucial role in exploring the internal logic of the decision process, making it understandable to humans [4], [12], [67].

The integration of XAI in cybersecurity has already witnessed significant success [10], [85], [91]. Numerous studies and implementations have demonstrated the efficacy of XAI in enhancing the interpretability and trustworthiness of malware detection systems. For example, in a seminal work Arp *et al.* [1] proposed a lightweight detection and explainable method called Drebin that leveraged a linear support vector machine to detect whether an unknown application is malware or benign, while also presenting to the user an overview of the important features that influenced the classification decision. Wu *et al.* [83] not only utilized a multi-layer perceptron (MLP) with an attention mechanism to detect malware but also pinpointed the key features and automatically produced natural language descriptions to interpret the core malicious behaviors of the flagged apps. Morcos *et al.* [56] employed the SHAP (Shapley Additive Explanations) method to interpret random forest models for identifying the most influential features for predicting malware or benign, along with quantifying their contributions for individual applications. Leveraging XAI approaches, Liu *et al.* [50] investigated why ML-based malware detection approaches perform so well under temporal inconsistency.

Despite these advancements, applying XAI in malware detection encounters several significant challenges. Firstly, there is a notable deficiency in evaluating how well these generated explanations align with the human-annotated ground truth of malware or the expectations of malware analysts and other stakeholders. Moreover, interpreting the explanation results often requires advanced domain-specific knowledge about malware, and thus XAI explanations are frequently inaccessible to non-technical audiences. Furthermore, the majority of XAI methods used in malware analysis are found to be unstable [21], [22], [45] and thus might leave the analysts confused when they query the system several times for the same sample. Additionally, in the malware detection domain clustering is an important method to identify malware families [5], [76], [88], but there is currently a scarcity of explainable clustering methods for malware family detection. Lastly, we note that deploying XAI in malware detection might pose additional risks, as it may introduce new vulnerabilities into malware detection pipelines.

In this paper, we systematize and discuss these challenges in more detail. We look at recent advancements in the XAI field in the AI literature and outline new research directions and opportunities that computer security and malware researchers might want to explore. We thus hope that this paper can help the community understand the current landscape of XAI techniques being applied to malware detection and identify promising research avenues.

XAI systematization of knowledge studies: XAI methods is a booming research area, and there have been several surveys and systematization of knowledge studies that summarized recent developments [14], [18], [26]. Particularly, Mohseni *et al.* [53] systematically studied design and evaluation techniques for XAI methods, with a focus on the different stakeholders involved. Gao *et al.* [24] systematically surveyed the explanation-guided

DOI 10.1109/EuroSPW61312.2024.00067

learning methods and discussed their performance evaluation criteria. Some recent surveys also focused on different AI subfields, such as deep learning [67], reinforcement learning [65], [80] or anomaly detection [47].

Moreover, several survey papers on XAI methods applied in the cybersecurity field appeared recently [10], [11], [59], [85], [91]. For example, Nadeem *et al.* [59] systematically studied the XAI methods applied for defensive and offensive cybersecurity tasks and discussed open research challenges. Vigano and Magazzeni [78] discussed the multi-faceted, multi-stakeholder nature of modern AI systems and propose the framework of *Explainable Security* (XSec) that guides the design of trustworthy XAI systems. Bhusal *et al.* [7] evaluated common XAI methods and discussed open research challenges in the security analytics domain. As one of the considered use cases, they experimented with a set of popular XAI methods applied to PDF malware detection, and provided quantitative and qualitative assessment results.

Finally, specifically in the malware detection field, a recent survey by Lin and Chang [49] reviewed the XAI methods applied for malware detection and proposed an interpretability score aggregating several important requirements from Molnar [55] that can be used to evaluate and compare XAI techniques.

In contrast to these works and drawing from their insights, our position paper aims to *summarize several open research challenges for XAI methods in application to malware detection*. To the best of our knowledge, these have not yet been discussed systematically, and our paper is a first step in this direction.

2. Alignment with Human Explanations

According to the systematic survey on explanation-guided learning by Gao *et al.* [24], XAI methods performance evaluation can be categorized into two pivotal dimensions: *alignment with human explanations* and *faithfulness to model predictions*. Alignment with human explanations focuses on whether human experts would consider the provided explanations correct and comprehensible. Faithfulness revolves around the concept of whether the model explanation remains true to the underlying model's reasoning, for example, under perturbations and when evaluating similar instances.

The alignment between the model's and human explanations is also called *correctness*. It focuses on the accuracy and perceptibility of the model explanations from a human perspective, delving into questions like "*How well does the model explanation align with the human explanation?*" and "*How well can humans comprehend the model explanation?*" [24].

These questions are far from trivial. For example, XAI methods often provide explanations in continuous values, representing the contribution or importance of each input feature to the model's prediction. For instance, for an Android malware detection model, an explainer might indicate that a feature `SEND_SMS` contributes 0.8 to detecting an app as malware, while another feature `UrlConnect()` contributes 0.2 to the same detection result. These continuous values allow for a subtle understanding of how different features influence the model's output. Yet, human explanations, especially when annotated for understanding, tend to be binary [24]. When humans annotate the ground truth, they might label features as either important or not important [15]. This binary approach simplifies the explanations but might lack the granularity provided by continuous values. The challenge arises when trying to align these two types of explanations. Continuous values from model explainers provide a gradient of importance across features, which does not directly map onto the binary categorization often used in human annotations. At the same time, it is not clear how to classify a feature importance value of 0.6 from a model explainer into binary terms. The importance threshold for this decision can be subjective and may not consistently align with human annotations. Translating continuous values into a format that can be compared or aligned with binary human annotations is thus needed.

Takeaway: XAI methods may output results that do not correspond to human annotations. Systematic approaches to reconciling malware explanations from a model and from a human analyst are needed.

2.1. Alignment Measurement Gaps

Substantial efforts have aimed to solve the alignment problem in the AI community. For instance, Doshi *et al.* [16] introduced simulatability and counterfactual simulation as metrics to assess interpretability. Concurrently, Hase *et al.* [27] discovered that LIME enhances the simulatability of models with tabular data. However, they noted that subjective evaluations of the explanations did not necessarily correlate with their practical utility. The study by Mohseni *et al.* [53] compared model explanations against human subjective assessments and evaluations based on ground-truth single-layer segmentation masks. Additional methods, such as mental model assessments, satisfaction, and trust evaluations via interviews and questionnaires, have been deployed to reconcile differences between model-generated and human-generated explanations [33], [54]. Furthermore, Liao *et al.* [48] developed an extensive XAI question bank for better capturing user requirements. However, as we discuss next, only limited work focuses on the explanation alignment in malware detection.

Evaluation methods for XAI techniques' alignment with human explanations broadly fall into three categories: case studies, user studies, and human annotation-based evaluations [24], [53]. A case study is an in-depth discussion of specific instances (an app or a malware family) where explanations are provided by the model. This is a common method to analyze and explain model-generated results in malware detection [1], [52], [63]. A user study is another qualitative evaluation method, involving participants (users) in evaluating the quality of explanations generated by the model. This can be achieved by, for example, creating a user interface that displays model explanations to human subjects, and the subjects rating the likelihood of the important features identified in the model explanations leading to a correct prediction of the underlying ground-truth label [33], [54]. For example, in the AI field, van der Waa *et al.* [77] compared rule-based and example-based XAI methods in the context of decision support in diabetes self-management. They discovered that rule-based explanations seem to yield a slightly better system understanding, while both rule- and example-based explanations are able to persuade users to follow the advice even when it is not correct. However, neither of the XAI methods improved the task performance for the users.

In the cybersecurity field, the study by Holder and Wang [28] applied document analysis and expert reviews for delving into stakeholder requirements for XAI systems deployed for cyber missions. Yet, while there have been several user studies with malware analysts [41], [84], [89], we are not aware of any systematic user studies involving explainable malware detection methods. The only study involving expert evaluation of XAI method results for malware detection was reported by Bhusal *et al.* [7], which had a very small sample size (n=1). The lack of user studies focusing on XAI techniques in the broader cybersecurity domain has also been noted by Nadeem *et al.* in their systematization-of-knowledge paper [59].

Many case studies and user studies assess whether a human understands the model-generated explanation, without evaluating the level of understanding. Instead, the human annotation-based evaluation methods can measure how well the human-annotated ground truth explanations are aligned with the model-generated explanations. In the AI community, Sen *et al.* [70] performed a quantitative comparative analysis (behavioral similarity) of machine attention maps created by deep learning models and human attention maps. Mohseni *et al.* [53] captured human annotations of salient features to create a human-grounded benchmark and investigate the relationship between subjective and objective evaluation of saliency explanations by comparing the benchmark with a binary feature mask ground truth (an objective measure) and user rating evaluations (a subjective measure). Atanasova *et al.* [2] compared the saliency scores assigned by the explainability techniques with human annotations of salient input regions to find relations between the model's performance and the agreement of its rationales with humans. Moreover, intersection over union (also called the Jaccard index) [44], mean average precision [17], precision, recall, and F1 are used to calculate the distance between ground truth and model-generated explanations [73].

However, currently, there is limited work focusing on establishing and measuring the explanation alignment in malware detection. The only study in this area is the XMal method proposed by Wu *et al.* [83] that estimated the alignment of synthesized natural language explanations with a ground truth on malware families.

Takeaway: There is a lack of evaluation metrics assessing the alignment of XAI-based malware explanations. While there are qualitative analysis methods such as case studies that showcase some explanations to demonstrate their comprehensibility to domain experts, they do not measure the level of understanding. It is necessary to develop metrics for evaluating the alignment, which can assess the level of the users' understanding and compare the comprehensibility of different XAI techniques.

2.2. Explanations for Non-Technical Audiences

As another aspect related to alignment, we should note that both expert and non-expert users of common XAI systems require usable explanations concerning their breadth and depth of domain knowledge [31]. Nadeem *et al.* [59] highlight the diversity of stakeholders in cybersecurity, each with distinct intents and expertise levels, interacting with the same ML models but for different purposes. These stakeholders require explanations at varying levels of detail and with different aims. For example, model users use explanations to gain insights into why an app was classified as malicious [51], and model designers apply explanations to find the causes of misclassification and ensure that the model employs meaningful features [6].

In the malware detection domain, XAI methods usually offer insights in the form of key features and rules that integrate, for example, system API calls. Such explanations, while valuable, demand a significant degree of domain-specific knowledge to be fully comprehensible. The process of interpreting these explanations can be particularly daunting for non-experts. For example, end-users may find it challenging to grasp the technical nuances of the explanations without a foundational understanding of malware and its manifestations. This gap underscores a critical barrier in the democratization of cybersecurity, where the benefits of advanced malware detection and explanation systems are limited by the technical proficiency required to interpret them.

To design a more comprehensible XAI system, Wu *et al.* [83] proposed a semantic matching of key features, which aims to contextualize explanation results within a framework understandable to various stakeholders. This approach represents a step forward by integrating expert knowledge into the system, thereby providing stakeholders with more accessible explanations. However, the prevalent trend in XAI techniques applied for malware detection still leans heavily towards technical outputs [49] that may not be readily accessible to non-technical audiences.

In the AI literature, as we mentioned, there have been studies on comparing human attention and annotations with the ones produced by XAI systems. For example, Reiter [68] summarized the natural language generation challenges for synthesizing usable explanations. Yet, as mentioned, so far there have been no studies with malware detection experts and malware detection system users to understand their needs and preferences for explanations.

Takeaway: To address this gap, the generated explanations need to be tailored to a specific purpose and audience, catering to the varying levels of comprehension of all stakeholders involved in malware detection. User studies with malware detection stakeholders can help to understand the requirements for automatically generated malware explanations.

3. Instability of Malware Explanations

As we mentioned, the second component of the performance evaluation of XAI methods according to Gao *et al.* [24] is *faithfulness to the ML model's predictions*. Faithfulness revolves around the concept of whether the model explanation remains true to the underlying model's reasoning, for example, under perturbations and when evaluating similar instances. Several criteria for assessing faithfulness have been proposed in the literature. For example, Yang *et al.* [87] summarized three significant properties from different perspectives, *i.e.*, generalizability, fidelity, and persuasibility. Ganz *et al.* [23] presented a framework for evaluating explanation methods on GNNs for vulnerability discovery and developed a set of criteria, including descriptive accuracy, structural robustness (whether explanation results change with perturbations), contrastivity, graph sparsity, stability (whether explanation results change in different runs), as well as efficiency (whether the explanation results are important to the decision making) for comparing graph explanations and linking them to properties of source code. Warnecke *et al.* [81] formulated conciseness, sparsity, completeness, and efficiency evaluation metrics to assess the performance of explanation methods.

In malware detection, there have been several studies of XAI methods' performance according to several faithfulness criteria. Fan *et al.* [21] assessed five well-known local and model-agnostic explanation approaches (LIME, Anchor, LORE, SHAP, and LEMNA), and Li and Gadyatskaya [45] evaluated five global explanation methods (SIRUS, deepRED, REM-D, ECLAIRE, and inTrees) for robustness, stability, and effectiveness. Warnecke *et al.* [82] introduced metrics such as the accuracy of explanations, completeness, efficiency, and robustness.

Stability is a property that measures the consistency of explanations generated for identical or similar instances in the data [79], and it is an important part of faithfulness. Since XAI methods attempt to provide insight into otherwise "black box" models, the provided explanations must be reliable. But, when the method is subject to randomness, there may be variations in the explanations, calling its reliability into question [66]. In the malware detection domain, stable explanation results are particularly crucial, because classification into malware families is based on similarity among the samples. If samples from the same family receive widely different explanations, malware researchers might miss the right family attribution or become confused [21]. Moreover, stable explanations support the system's reliability and the model users' and designers' trust, and are instrumental in debugging and improving the system, enabling developers to address deficiencies effectively. In essence, we can argue that the stability of explanation results is not just a technical requirement but a foundational aspect of user trust and the overall efficacy of malware detection systems.

Yet, many popular XAI methods are inherently unstable. The study by Li and Gadyatskaya [45] has shown that several established global XAI methods have stability scores equal to zero when applied to malware detection, while Fan *et al.* [21] found that stability scores for the popular local XAI methods LEMNA, LORE, and Anchor are below 0.5.

This instability, a consequence of the plurality of explanations produced by the same model on the same sample, is known as *the disagreement problem* in the AI community [25], [34], [58], [60]. It follows from the notion of *the Rashomon effect* for ML and XAI methods, which describes the fact that different statistical models can work equally well when fitting the same data [9], [40], [58]. In a user study, Krishna *et al.* [34] showed that practitioners are indeed aware that different XAI methods do not agree in their explanations, and they often resort to arbitrary choices, e.g., just rely on their favorite method. It is thus important that the malware detection community pays more attention to this issue. A possible solution for XAI-enabled malware detection methods that will not be confusing to analysts can lie in combining several explainers, as proposed by, e.g., Pirie *et al.* [64]. Another direction to explore is the Functional Decomposition tree approach proposed by Laberge *et al.* [38].

Takeaway: XAI methods for malware detection have high stability requirements, while many established XAI methods are unstable. Further research into more stable explanation methods is required.

4. No Studies of XAI for Malware Clustering

Supervised learning techniques, where the model learns from labeled data to predict outcomes, have been the center of attention in the AI community as well as in malware detection. Correspondingly, the research of XAI methods based on supervised learning and applying these XAI methods in malware detection has also flourished. Yet, there are also many established approaches for unsupervised learning methods for malware detection and categorization [5], [13], [20], [20]. Clustering algorithms segment data into groups based on similarities without prior labeling, offering a unique perspective in identifying novel or unknown malware types [5], [76], [88]. It would thus be useful for security analysts to receive explanations as to why certain samples are grouped into a family. Considering the explanation approaches for unsupervised learning methods, we observe that there are several works on explaining the results of unsupervised models in the AI community and the absence of research focusing on the explainability of clustering in malware detection.

In the AI community, Kauffmann *et al.* [32] proposed a framework that can explain cluster assignments in terms of input features that have contributed to the cluster assignments by rewriting clustering models as neural networks. Moshkovitz *et al.* [57] used a small decision tree to partition a data set into clusters, which can characterize clusters straightforwardly. Based on this research, Laber *et al.* [37] proposed a simple greedy algorithm for building explainable clustering for the k-means method, and Eduardo *et al.* [36] constructed shallow decision trees — i.e., trees whose leaves are not very deep, which translates into clusters that are defined by a small number of attributes. Lawless *et al.* [39] provided a new approach that clusters data points and constructs polytopes around the discovered clusters to explain them.

In malware detection, research focusing on the explainability of clustering is, to the best of our knowledge, unexplored, which limits the understanding of how these models discern patterns and relationships in malware samples. Demystifying the operation of clustering algorithms and elucidating the features that influence each cluster can help identify new malware strains and provide evidence to support the decision. Addressing this challenge is thus very important for advancing the field of XAI in malware detection, ensuring that both supervised and unsupervised learning methods are transparent, interpretable, and, consequently, more trustworthy and effective.

Takeaway: There have been no studies of explainable clustering methods for malware detection. New research in this area can help elucidate novel patterns in emerging malware strains. Additionally, it will be interesting to compare XAI methods for unsupervised learning-based malware detection methods based on their alignment and faithfulness, as previous studies focused only on XAI techniques for supervised learning.

5. Attack Risks on Malware Explanations

While model explanations are designed to improve transparency, they inadvertently create new vulnerabilities, and interpretability is potentially susceptible to malicious manipulations [90]. As reported by Nadeem *et al.* [59], adversaries might exploit the insights derived from these explanations to initiate new attacks. Notably, state-of-the-art XAI methods such as LIME and SHAP are not impervious to such threats [74]. Zhang *et al.* [90] demonstrated this by generating adversarial inputs to compromise explainable deep learning systems, impacting not just the target deep neural networks but also the associated interpretative models, thereby casting doubts on the reliability of current XAI methods.

Further emphasizing this vulnerability, Kuppa *et al.* [35] employed XAI techniques to undermine the classifiers' confidentiality and robustness using counterfactual explanations. They utilized a dataset of 30,120 malware samples to test evasion tactics encompassing explanation-based poisoning and evasion attacks. The study has shown that XAI methods can be adversarially exploited and also shed light on privacy attacks [19] encompassing explanation-based model extraction, and membership inference attacks [29] validated using leaked password datasets and network traffic data.

In response to these threats, the community needs to devise defense strategies to counteract explanation-aware threat scenarios and safeguard XAI-enabled malware detection systems. The literature extensively discusses countermeasures against prediction-only attacks, i.e., when the small perturbation in data will mislead the model to predict wrong results [8], [46], [69]. Specifically in the malware detection realm, we can mention, for example, [42], [43], [69], [86]. For instance, Sun *et al.* [75] and Severi *et al.* [71] demonstrated that explainability-guided adversarial modifications significantly worsened the detection performance of malware classifiers, and they proposed methods to improve classifier robustness by targeted feature selection and applying anomaly detection techniques for identifying perturbed intruders. Still, strategies to counter explanation-aware attacks remain markedly underdeveloped [62]. For instance, robustness improvement methods like AMM [75] require the ML model designers to have in-depth expert knowledge about which features can be feasibly added or modified in malware samples without damaging the payload. This may not be feasible for all stakeholders, especially considering that advanced malware writers constantly look for new ways to implement malicious functionality [30], [61], [72].

Takeaway: Strengthening the explainability of XAI models against explanation-aware threats presents a challenge that the AI and security communities should try to address.

6. Conclusion

In this paper, we discuss the multifaceted challenges associated with the current research on applying XAI in malware detection and analysis. We reveal that there is a lack of methods and studies to assess the XAI techniques' alignment with human explanations, with a particular lack of support for non-technical audiences, underscoring a significant gap in accessibility. Moreover, we observe a prevalent instability in popular XAI methods applied to interpret the malware detection process, which might result in a lack of trust from the users. The absence of XAI techniques for clustering methods tailored for malware detection further complicates the interpretability of results. We also highlight the inherent risks associated with deploying XAI in malware detection, potentially exposing the system to sophisticated explanation-aware attack vectors.

Acknowledgements. This research was partially supported by the Chinese Scholarship Council (CSC). We warmly thank the anonymous reviewers and our shepherd for their constructive feedback and insightful comments that helped us to improve this paper.

References

[1] Daniel Arp, Michael Spreitzenbarth, Malte Hübner, Hugo Gascon, and Konrad Rieck. Drebin: Effective and explainable detection of Android malware in your pocket. In *Symposium on Network and Distributed System Security (NDSS)*, 02 2014.

[2] Pepa Atanasova, Jakob Grue Simonsen, Christina Lioma, and Isabelle Augenstein. A diagnostic study of explainability techniques for text classification. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3256–3274, Online, November 2020. Association for Computational Linguistics.

[3] AV-ATLAS. Malware statistics, 2024. https://portal.av-atlas.org/malware/statistics.

[4] Xiao Bai, Xiang Wang, Xianglong Liu, Qiang Liu, Jingkuan Song, Nicu Sebe, and Been Kim. Explainable deep learning for efficient and robust pattern recognition: A survey of recent developments. *Pattern Recognition*, 120:108102, 2021.

[5] Ulrich Bayer, Paolo Milani Comparetti, Clemens Hlauschek, Christopher Kruegel, and Engin Kirda. Scalable, behavior-based malware clustering. In *NDSS*, volume 9, pages 8–11, 2009.

[6] Franziska Becker, Arthur Drichel, Christoph Müller, and Thomas Ertl. Interpretable visualizations of deep neural networks for domain generation algorithm detection. In *2020 IEEE Symposium on Visualization for Cyber Security (VizSec)*, pages 25–29, 2020.

[7] Dipkamal Bhusal, Rosalyn Shin, Ajay Ashok Shewale, Monish Kumar Manikya Veerabhadran, Michael Clifford, Sara Rampazzi, and Nidhi Rastogi. SoK: Modeling explainability in security analytics for interpretability, trustworthiness, and usability. In *Proceedings of the 18th International Conference on Availability, Reliability and Security*, pages 1–12, 2023.

[8] Panagiotis Bountakas, Apostolis Zarras, Alexios Lekidis, and Christos Xenakis. Defense strategies for adversarial machine learning: A survey. *Computer Science Review*, 49:100573, 2023.

[9] Leo Breiman. Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical science*, 16(3):199–231, 2001.

[10] Nicola Capuano, Giuseppe Fenza, Vincenzo Loia, and Claudio Stanzione. Explainable artificial intelligence in cybersecurity: A survey. *IEEE Access*, 10:93575–93600, 2022.

[11] Fabien Charmet, Harry Chandra Tanuwidjaja, Solayman Ayoubi, Pierre-François Gimenez, Yufei Han, Houda Jmila, Gregory Blanc, Takeshi Takahashi, and Zonghua Zhang. Explainable artificial intelligence for cybersecurity: a literature survey. *Annals of Telecommunications*, 77(11):789–812, 2022.

[12] Valerie Chen, Jeffrey Li, Joon Sik Kim, Gregory Plumb, and Ameet Talwalkar. Interpretable machine learning: Moving from mythos to diagnostics. *Communications of the ACM*, 65(8):43–50, 2022.

[13] Khanh Huu The Dam, Thomas Given-Wilson, and Axel Legay. Unsupervised behavioural mining and clustering for malware family identification. In *Proceedings of the 36th Annual ACM Symposium on Applied Computing*, pages 374–383, 2021.

[14] Arun Das and Paul Rad. Opportunities and challenges in explainable artificial intelligence (XAI): A survey. *arXiv preprint arXiv:2006.11371*, 2020.

[15] Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C. Wallace. ERASER: A benchmark to evaluate rationalized NLP models. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4443–4458, Online, July 2020. Association for Computational Linguistics.

[16] Finale Doshi-Velez and Been Kim. Towards a rigorous science of interpretable machine learning. *arXiv: Machine Learning*, 2017.

[17] Mengnan Du, Ninghao Liu, Qingquan Song, and Xia Hu. Towards explanation of dnn-based prediction with guided feature inversion. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '18, page 1358–1367, New York, NY, USA, 2018. Association for Computing Machinery.

[18] Rudresh Dwivedi, Devam Dave, Het Naik, Smiti Singhal, Rana Omer, Pankesh Patel, Bin Qian, Zhenyu Wen, Tejal Shah, and Graham Morgan. Explainable AI (XAI): Core ideas, techniques, and solutions. *ACM Computing Surveys*, 55(9):1–33, 2023.

[19] Cynthia Dwork, Adam Smith, Thomas Steinke, and Jonathan Ullman. Exposed! a survey of attacks on private data. *Annual Review of Statistics and Its Application*, 4:61–84, 2017.

[20] Ming Fan, Xiapu Luo, Jun Liu, Meng Wang, Chunyin Nong, Qinghua Zheng, and Ting Liu. Graph embedding based familial analysis of Android malware using unsupervised learning. In *2019 IEEE/ACM 41st International Conference on Software Engineering (ICSE)*, pages 771–782. IEEE, 2019.

[21] Ming Fan, Wenying Wei, Xiaofei Xie, Yang Liu, Xiaohong Guan, and Ting Liu. Can we trust your explanations? Sanity checks for interpreters in Android malware analysis. *IEEE Transactions on Information Forensics and Security*, 16:838–853, 2020.

[22] Antonio Galli, Valerio La Gatta, Vincenzo Moscato, Marco Postiglione, and Giancarlo Sperlì. Explainability in AI-based behavioral malware detection systems. *Computers & Security*, 141:103842, 2024.

[23] Tom Ganz, Martin Härterich, Alexander Warnecke, and Konrad Rieck. Explaining graph neural networks for vulnerability discovery. In *Proceedings of the 14th ACM Workshop on Artificial Intelligence and Security*, AISec '21, page 145–156, New York, NY, USA, 2021. Association for Computing Machinery.

[24] Yuyang Gao, Siyi Gu, Junji Jiang, Sungsoo Ray Hong, Dazhou Yu, and Liang Zhao. Going beyond XAI: A systematic survey for explanation-guided learning. *ACM Computing Surveys*, feb 2024.

[25] Sofie Goethals, David Martens, and Theodoros Evgeniou. Manipulation risks in explainable AI: The implications of the disagreement problem. *arXiv preprint arXiv:2306.13885*, 2023.

[26] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. A survey of methods for explaining black box models. *ACM Comput. Surv.*, 51(5), aug 2018.

[27] Peter Hase and Mohit Bansal. Evaluating explainable AI: Which algorithmic explanations help users predict model behavior? In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5540–5552, Online, July 2020. Association for Computational Linguistics.

[28] Eric Holder and Ning Wang. Explainable artificial intelligence (xai) interactively working with humans as a junior cyber analyst. *Human-Intelligent Systems Integration*, 3(2):139–153, 2021.

[29] Hongsheng Hu, Zoran Salcic, Lichao Sun, Gillian Dobbie, Philip S. Yu, and Xuyun Zhang. Membership inference attacks on machine learning: A survey. *ACM Comput. Surv.*, 54(11s), sep 2022.

[30] Roberto Jordaney, Kumar Sharad, Santanu K Dash, Zhi Wang, Davide Papini, Ilia Nouretdinov, and Lorenzo Cavallaro. Transcend: Detecting concept drift in malware classification models. In *26th USENIX security symposium (USENIX security 17)*, pages 625–642, 2017.

[31] Mladan Jovanovic and Mia Schmitz. Explainability as a user requirement for artificial intelligence systems. *Computer*, 55(2):90–94, 2022.

[32] Jacob Kauffmann, Malte Esders, Lukas Ruff, Grégoire Montavon, Wojciech Samek, and Klaus-Robert Müller. From clustering to cluster explanations via neural networks. *IEEE Transactions on Neural Networks and Learning Systems*, 2022.

[33] Harmanpreet Kaur, Harsha Nori, Samuel Jenkins, Rich Caruana, Hanna Wallach, and Jennifer Wortman Vaughan. Interpreting interpretability: Understanding data scientists' use of interpretability tools for machine learning. In *Proceedings of the 2020 CHI conference on human factors in computing systems*, pages 1–14, 2020.

[34] Satyapriya Krishna, Tessa Han, Alex Gu, Javin Pombra, Shahin Jabbari, Steven Wu, and Himabindu Lakkaraju. The disagreement problem in explainable machine learning: A practitioner's perspective. *arXiv preprint arXiv:2202.01602*, 2022.

[35] Aditya Kuppa and Nhien-An Le-Khac. Adversarial XAI methods in cybersecurity. *IEEE Transactions on Information Forensics and Security*, 16:4924–4938, 2021.

[36] Eduardo Laber, Lucas Murtinho, and Felipe Oliveira. Shallow decision trees for explainable k-means clustering. *Pattern Recognition*, 137:109239, 2023.

[37] Eduardo S Laber and Lucas Murtinho. On the price of explainability for some clustering problems. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 5915–5925. PMLR, 18–24 Jul 2021.

[38] Gabriel Laberge, Yann Batiste Pequignot, Mario Marchand, and Foutse Khomh. Tackling the XAI disagreement problem with regional explanations. In *International Conference on Artificial Intelligence and Statistics*, pages 2017–2025. PMLR, 2024.

[39] Connor Lawless, Jayant Kalagnanam, Lam M Nguyen, Dzung Phan, and Chandra Reddy. Interpretable clustering via multi-polytope machines. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(7):7309–7316, Jun. 2022.

[40] Anastasia Leventi-Peetz and Kai Weber. Rashomon effect and consistency in explainable artificial intelligence (XAI). In *Proceedings of the Future Technologies Conference*, pages 796–808. Springer, 2022.

[41] Fanny Lalonde Lévesque, Sonia Chiasson, Anil Somayaji, and José M Fernandez. Technological and human factors of malware attacks: A computer security clinical trial approach. *ACM Transactions on Privacy and Security (TOPS)*, 21(4):1–30, 2018.

[42] Deqiang Li and Qianmu Li. Adversarial deep ensemble: Evasion attacks and defenses for malware detection. *IEEE Transactions on Information Forensics and Security*, 15:3886–3900, 2020.

[43] Deqiang Li, Qianmu Li, Yanfang Ye, and Shouhuai Xu. Arms race in adversarial malware detection: A survey. *ACM Computing Surveys (CSUR)*, 55(1):1–35, 2021.

[44] Kunpeng Li, Ziyan Wu, Kuan-Chuan Peng, Jan Ernst, and Yun Fu. Tell me where to look: Guided attention inference network. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9215–9223, 2018.

[45] Rui Li and Olga Gadyatskaya. Evaluating rule-based global XAI malware detection methods. In *International Conference on Network and System Security*, pages 3–22. Springer, 2023.

[46] Yao Li, Minhao Cheng, Cho-Jui Hsieh, and Thomas CM Lee. A review of adversarial attack and defense for classification methods. *The American Statistician*, 76(4):329–345, 2022.

[47] Zhong Li, Yuxuan Zhu, and Matthijs Van Leeuwen. A survey on explainable anomaly detection. *ACM Transactions on Knowledge Discovery from Data*, 18(1):1–54, 2023.

[48] Qingzi Vera Liao, Dan Gruen, and Sarah Miller. Questioning the ai: Informing design practices for explainable AI user experiences. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 2020.

[49] Yuzhou Lin and Xiaolin Chang. Towards interpreting ML-based automated malware detection models: A survey. *arXiv preprint arXiv:2101.06232*, 2021.

[50] Yue Liu, Chakkrit Tantithamthavorn, Li Li, and Yepang Liu. Explainable AI for Android malware detection: Towards understanding why the models perform so well? In *2022 IEEE 33rd International Symposium on Software Reliability Engineering (ISSRE)*, pages 169–180, 2022.

[51] Sherin Mary. *Explainable Artificial Intelligence Applications in NLP, Biomedical, and Malware Classification: A Literature Review*, pages 1269–1292. 07 2019.

[52] Marco Melis, Davide Maiorca, Battista Biggio, Giorgio Giacinto, and Fabio Roli. Explaining black-box Android malware detection. In *2018 26th european signal processing conference (EUSIPCO)*, pages 524–528. IEEE, 2018.

[53] Sina Mohseni, Jeremy E Block, and Eric Ragan. Quantitative evaluation of machine learning explanations: A human-grounded benchmark. In *26th International Conference on Intelligent User Interfaces*, IUI '21, page 22–31, New York, NY, USA, 2021. Association for Computing Machinery.

[54] Sina Mohseni, Niloofar Zarei, and Eric D. Ragan. A multidisciplinary survey and framework for design and evaluation of explainable AI systems. *ACM Trans. Interact. Intell. Syst.*, 11(3–4), sep 2021.

[55] Christoph Molnar. *Interpretable machine learning*. Lulu.com, 2020.

[56] Martina Morcos, Hussam Al Hamadi, Ernesto Damiani, Sivaprasad Nandyala, and Brian McGillion. A surrogate-based technique for Android malware detectors' explainability. In *2022 18th International Conference on Wireless and Mobile Computing, Networking and Communications (WiMob)*, pages 112–117. IEEE, 2022.

[57] Michal Moshkovitz, Sanjoy Dasgupta, Cyrus Rashtchian, and Nave Frost. Explainable k-means and k-medians clustering. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 7055–7065. PMLR, 13–18 Jul 2020.

[58] Sebastian Müller, Vanessa Toborek, Katharina Beckh, Matthias Jakobs, Christian Bauckhage, and Pascal Welke. An empirical evaluation of the Rashomon effect in explainable machine learning. In *Proc. of ECML/PKDD*, pages 462–478. Springer, 2023.

[59] Azqa Nadeem, Daniël Vos, Clinton Cao, Luca Pajola, Simon Dieck, Robert Baumgartner, and Sicco Verwer. SoK: Explainable machine learning for computer security applications. In *IEEE European Symposium on Security and Privacy*, pages 221–240, 2023.

[60] Michael Neely, Stefan F Schouten, Maurits JR Bleeker, and Ana Lucic. Order in the court: Explainable AI methods prone to disagreement. *arXiv preprint arXiv:2105.03287*, 2021.

[61] Lily Hay Newman. How Android fought an epic botnet —- and won, 2019.

[62] Maximilian Noppel and Christian Wressnegger. SoK: Explainable machine learning in adversarial environments. In *2024 IEEE Symposium on Security and Privacy (SP)*, pages 21–21. IEEE Computer Society, 2023.

[63] Lukas Pirch, Alexander Warnecke, Christian Wressnegger, and Konrad Rieck. Tagvet: Vetting malware tags using explainable machine learning. In *Proceedings of the 14th European Workshop on Systems Security*, pages 34–40, 2021.

[64] Craig Pirie, Nirmalie Wiratunga, Anjana Wijekoon, and Carlos Francisco Moreno-Garcia. AGREE: A feature attribution aggregation framework to address explainer disagreements with alignment metrics. CEUR Workshop Proceedings, 2023.

[65] Erika Puiutta and Eric Veith. Explainable reinforcement learning: A survey. In *International cross-domain conference for machine learning and knowledge extraction*, pages 77–95. Springer, 2020.

[66] Guidotti R. and Ruggieri S. On the stability of interpretable models. In *IJCNN 2019 - International Joint Conference on Neural Networks (IJCNN), Budapest, Hungary, 14-19 July, 2019*, 2019.

[67] Gabrielle Ras, Ning Xie, Marcel Van Gerven, and Derek Doran. Explainable deep learning: A field guide for the uninitiated. *Journal of Artificial Intelligence Research*, 73:329–396, 2022.

[68] Ehud Reiter. Natural language generation challenges for Explainable AI. In Jose M. Alonso and Alejandro Catala, editors, *Proceedings of the 1st Workshop on Interactive Natural Language Technology for Explainable Artificial Intelligence (NL4XAI 2019)*, pages 3–7. Association for Computational Linguistics, 2019.

[69] Ishai Rosenberg, Asaf Shabtai, Yuval Elovici, and Lior Rokach. Adversarial machine learning attacks and defense methods in the cyber security domain. *ACM Computing Surveys (CSUR)*, 54(5):1–36, 2021.

[70] Cansu Sen, Thomas Hartvigsen, Biao Yin, Xiangnan Kong, and Elke Rundensteiner. Human attention maps for text classification: Do humans and neural networks focus on the same words? In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4596–4608, Online, July 2020. Association for Computational Linguistics.

[71] Giorgio Severi, Jim Meyer, Scott Coull, and Alina Oprea. Explanation-guided backdoor poisoning attacks against malware classifiers. In *30th USENIX security symposium (USENIX security 21)*, pages 1487–1504, 2021.

[72] Anshuman Singh, Andrew Walenstein, and Arun Lakhotia. Tracking concept drift in malware families. In *Proceedings of the 5th ACM workshop on Security and artificial intelligence*, pages 81–92, 2012.

[73] Xuelin Situ, Ingrid Zukerman, Cecile Paris, Sameen Maruf, and Gholamreza Haffari. Learning to explain: Generating stable explanations fast. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5340–5355, Online, August 2021. Association for Computational Linguistics.

[74] Dylan Slack, Sophie Hilgard, Emily Jia, Sameer Singh, and Himabindu Lakkaraju. Fooling LIME and SHAP: Adversarial attacks on post hoc explanation methods. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, AIES '20, page 180–186, New York, NY, USA, 2020. Association for Computing Machinery.

[75] Ruoxi Sun, Minhui Xue, Gareth Tyson, Tian Dong, Shaofeng Li, Shuo Wang, Haojin Zhu, Seyit Camtepe, and Surya Nepal. Mate! Are you really aware? An explainability-guided testing framework for robustness of malware detectors. In *Proceedings of the 31st ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, ESEC/FSE 2023, page 1573–1585, New York, NY, USA, 2023. Association for Computing Machinery.

[76] Daniele Ucci, Leonardo Aniello, and Roberto Baldoni. Survey of machine learning techniques for malware analysis. *Computers & Security*, 81:123–147, 2019.

[77] Jasper van der Waa, Elisabeth Nieuwburg, Anita Cremers, and Mark Neerincx. Evaluating XAI: A comparison of rule-based and example-based explanations. *Artificial Intelligence*, 291:103404, 2021.

[78] L. Vigano and D. Magazzeni. Explainable security. In *2020 IEEE European Symposium on Security and Privacy Workshops (EuroS&PW)*, pages 293–300. IEEE, 2020.

[79] Giorgio Visani, Enrico Bagli, Federico Chesani, Alessandro Poluzzi, and Davide Capuzzo. Statistical stability indices for LIME: Obtaining reliable explanations for machine learning models. *Journal of the Operational Research Society*, 73(1):91–101, 2022.

[80] George A Vouros. Explainable deep reinforcement learning: State of the art and challenges. *ACM Computing Surveys*, 55(5):1–39, 2022.

[81] Alexander Warnecke, Daniel Arp, Christian Wressnegger, and Konrad Rieck. Don't paint it black: White-box explanations for deep learning in computer security. *CoRR*, 2019.

[82] Alexander Warnecke, Daniel Arp, Christian Wressnegger, and Konrad Rieck. Evaluating explanation methods for deep learning in security. In *2020 IEEE european symposium on security and privacy (EuroS&P)*, pages 158–174. IEEE, 2020.

[83] Bozhi Wu, Sen Chen, Cuiyun Gao, Lingling Fan, Yang Liu, Weiping Wen, and Michael R. Lyu. Why an Android app is classified as malware: Toward malware classification interpretation. *ACM Transactions on Software Engineering and Methodology*, 30(2), 2021.

[84] Khaled Yakdan, Sergej Dechand, Elmar Gerhards-Padilla, and Matthew Smith. Helping Johnny to analyze malware: A usability-optimized decompiler and malware analysis user study. In *2016 IEEE Symposium on Security and Privacy (SP)*, pages 158–177. IEEE, 2016.

[85] Feixue Yan, Sheng Wen, Surya Nepal, Cecile Paris, and Yang Xiang. Explainable machine learning in cybersecurity: A survey. *International Journal of Intelligent Systems*, 37(12):12305–12334, 2022.

[86] Senming Yan, Jing Ren, Wei Wang, Limin Sun, Wei Zhang, and Quan Yu. A survey of adversarial attack and defense methods for malware classification in cyber security. *IEEE Communications Surveys & Tutorials*, 25(1):467–496, 2022.

[87] Fan Yang, Mengnan Du, and Xia Hu. Evaluating explanation without ground truth in interpretable machine learning. *ArXiv*, abs/1907.06831, 2019.

[88] Yanfang Ye, Tao Li, Donald Adjeroh, and S Sitharama Iyengar. A survey on malware detection using data mining techniques. *ACM Computing Surveys (CSUR)*, 50(3):1–40, 2017.

[89] Miuyin Yong Wong, Matthew Landen, Manos Antonakakis, Douglas M Blough, Elissa M Redmiles, and Mustaque Ahamad. An inside look into the practice of malware analysis. In *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security*, pages 3053–3069, 2021.

[90] Xinyang Zhang, Ningfei Wang, Hua Shen, Shouling Ji, Xiapu Luo, and Ting Wang. Interpretable deep learning under fire. In *Proceedings of the 29th USENIX Conference on Security Symposium*, SEC'20, USA, 2020. USENIX Association.

[91] Zhibo Zhang, Hussam Al Hamadi, Ernesto Damiani, Chan Yeob Yeun, and Fatma Taher. Explainable artificial intelligence applications in cyber security: State-of-the-art in research. *IEEE Access*, 10:93104–93139, 2022.