**DOI: 10.5281/zenodo.10669844**

# Exploring Explainable AI, Security and Beyond : A Comprehensive Review

Nikita Niteen
Amal Jyothi College of Engineering
Kanjirapally, India
nikitaniteen@cs.ajce.in

Simy Mary Kurian
Amal Jyothi College of Engineering
Kanjirapally, India
simymarykurian@amaljyothi.ac.in

*Abstract*—**The paper dives into the transformation of security with the integration of machine learning (ML) and the associated challenges. It highlights how AI's incorporation in network security brings both promise and complexities, emphasizing the need to align perceived benefits with actual capabilities. Explainable AI (XAI) emerges as a crucial tool, offering transparency despite facing ongoing challenges and necessitating continual advancements. The pursuit of Explainable AI (XAI) and tools like AI Explainability 360 demonstrate strengths but grapple with understanding and methodological gaps. Specific techniques such as LIME, GNNEXPLAINER, and object recognition in Deep Reinforcement Learning show promise but encounter challenges like scalability and adaptability. Across these domains, understanding the multifaceted landscape becomes pivotal for leveraging's potential while addressing critical challenges in security and explainability.**

*Index Terms*—**Machine Learning (ML), Big Data Frameworks, Deep Learning, Explainable AI (XAI), Reinforcement Learning, Natural Language Processing (NLP).**

## *I. INTRODUCTION*

Machine learning's integration into security domains has revolutionized threat detection and response capabilities, showcasing remarkable advantages in enhancing security processes. However, this fusion is not without its challenges, posing critical limitations that demand attention. Concurrently, the amalgamation of diverse technologies, including ML algorithms and big data frameworks, fortifies security measures. Understanding the landscape of advantages, limitations, and the technological spectrum becomes crucial in comprehending the pivotal role of machine learning in bolstering security measuresThe application of Artificial Intelligence (AI) in network security reflects both the potentials and complexities of incorporating this transformative technology. While emphasizing AI's adaptability in tackling evolving threats, the text underscores the need to bridge the gap between perceived benefits and actual capabilities. Furthermore, it addresses the pitfalls of hastening adoption without a comprehensive understanding, reflecting the ongoing assimilation of AI in cyber security. Explainable Artificial Intelligence (XAI) stands as a beacon in the realm of AI decision-making, offering transparency and comprehensibility in complex AI models. Despite its commendable attributes, challenges like conceptual ambiguity and susceptibility to adversarial attacks persist, underscoring the need for continuous advancements. The integration of XAI into diverse domains, including IoT applications, showcases promise but necessitates further security enhancements and tailored interfaces for comprehensive explanations. As the pursuit of Explainable AI (xAI) gains momentum, a panorama of its advantages and limitations unfolds. xAI excels in elucidating opaque AI models, yet faces limitations in achieving comprehensive understanding and clarity in debates surrounding its role. Amidst this landscape, advanced AI technologies, particularly deep neural networks, drive xAI research, urging the need for sociotechnical perspectives in addressing AI explainability. The AI Explainability 360 toolkit emerges as a comprehensive repository of explanation methods, offering diverse solutions for users across domains. While its structured taxonomy simplifies navigation, the toolkit grapples with gaps in certain explanation categories and potential framework dependence. Its technological foundation, featuring advanced algorithms and educational resources, significantly augments accessibility and understanding in explainable AI. LIME, a unique explanation technique, showcases significant strengths in trust-building through interpretable predictions. However, its focus on sparse linear models and challenges in refining image- based explanations highlight areas for improvement. Leveraging local model learning and optimization techniques, LIME excels in generating accurate and real-world applicable explanations. GNNEXPLAINER presents a breakthrough in interpretability for GNN-based models, offering insights into crucial graph pathways and node features. While addressing potential complexity issues in larger graphs, it leverages GNN architectures and optimization tasks for accurate and informative explanations. The integration of object recognition into Deep Reinforcement Learning models augments system transparency but introduces scalability concerns. By leveraging DRL frameworks and object recognition techniques, this integration holds promise for enhancing system transparency. The Generative Explanation Framework (GEF) revolutionizes NLP by offering fine-grained explanations aligned with classification

**DOI: 10.5281/zenodo.10669844**

decisions. However, its adaptability challenges highlight the need for seamless integration across a broader spectrum of NLP tasks. GEF technological foundation leverages tailored frameworks and innovative training methods to enhance model interpretability. The landscape of AI explainability, ranging from XAI to specialized toolkits, presents a diverse array of dvancements, limitations, and technological innovations. Understanding this multifaceted domain becomes pivotal in harnessing AIs full potential while addressing critical challenges. These introductions provide a glimpse into the multifaceted landscape of AI integration, explainability, and security, offering insights into their potentials, limitations, and technological foundations.

## *II. RELATED WORKS.*

### *A. SoK: Applying Machine Learning in Security - A Survey*

Integrating machine learning (ML) into security domains presents a range of advantages. ML algorithms, like neural networks and decision trees, significantly enhance threat detection capabilities by identifying intricate patterns within large datasets. This technology facilitates the automation of security processes, reducing manual efforts and minimizing false positives by prioritizing real threats based on historical data. Moreover, ML frameworks equipped with big data technologies offer adaptability to evolving threats, allowing security measures to adjust dynamically and respond effectively to changing patterns. However, this integration also faces several limitations. MLs dependency on historical data might hinder its ability to swiftly adapt to rapidly evolving threats, potentially leaving systems ill- prepared to recognize novel attack patterns lacking sufficient historical context. Theres also a concern regarding publication bias in academic literature, where favoritism toward specific high-ranked conferences might limit exposure to diverse perspectives and innovative approaches, potentially overlooking emerging methodologies. Additionally, acquiring substantial labeled data for training ML models poses a challenge, especially in security domains where labeled data might be scarce, sensitive, or expensive to obtain. The integration leverages a spectrum of technologies, including diverse ML algorithms, big data frameworks such as Apache Spark and Hadoop, and deep learning libraries like TensorFlow and PyTorch. These technologies collectively fortify security measures by enabling advanced pattern recognition, large-scale data processing, and implementation of sophisticated learning models

### *B. Applications of Artificial Intelligence (AI) to Network Security*

The passage emphasizes the complexities and potentials of employing Artificial Intelligence (AI) in network security. It underscores the challenges, including evolving attack complexities and the need to bridge the gap in understanding AI's actual capabilities versus perceived benefits. Additionally, it notes the rush to adopt buzzwords without a clear comprehension of their implications, reflecting the ongoing assimilation of AI in cybersecurity. On the positive side, it highlights the strengths of AI, particularly Machine Learning (ML), in adapting to evolving threats, enabling real-time detection, and reducing dependency on human intervention. It accentuates the paradigm shift from rule-based to reactive real-time detection methods empowered by AI and ML, offering promising insights into proactive threat identification. The text mentions the practicality of supervised ML in addressing specific security issues and the potential of unsupervised ML to reduce human involvement further. The integration of AI with data science techniques is also underscored, emphasizing the need for enhanced data visualization, contextual understanding, and tighter integration within existing security frameworks. In terms of technologies, the discussion alludes to the utilization of various ML algorithms, deep learning frameworks, big data processing tools, data visualization libraries, security analytics platforms, and cloud computing services, showcasing the multifaceted technological landscape integral to AI's integration in network security.

### *C . XAI for Cybersecurity: State of the Art, Challenges, Open Issues and Future Directions.*

Explainable Artificial Intelligence (XAI) serves as a solution to the opaqueness of AI models, offering transparency and understandable interpretations. Its versatile application spans critical domains like healthcare, military, energy, finance, and industry, enhancing comprehension in complex decision-making systems. Conceptual ambiguity poses challenges, as XAI encounters dilemmas regarding loyalty to models or data generation processes, leading to misunderstandings and debates in interpreting AI decisions. Additionally, the absence of standardized measurement metrics hampers assessing the quality of interpretability, impacting universal understanding. Vulnerabilities to adversarial attacks raise concerns about trust and security, as some XAI models are susceptible to manipulation by adversaries. Specialized methodologies tailored to models enhance fidelity and completeness of explanations, delving deeper into model architectures. Automation akin to ML tasks simplifies explanations, improves feature selection, and bolsters security, particularly crucial in IoT applications. Leveraging federated learning in IoT networks allows fog and cloud servers to explain models locally and globally, contributing to distributed AI decision-making. Explainable AI (XAI) addresses AI's opacity but faces challenges like conceptual ambiguity, lack of measurement metrics, and susceptibility to attacks. Its integration into IoT applications shows promise but necessitates improvements in security and tailored interfaces for comprehensive explanations. Successful deployment in AI-driven IoT applications relies on understanding these benefits, limitations, and technological considerations.

**DOI: 10.5281/zenodo.10669844**

### *D. Explainable Artificial Intelligence (XAI) for Internet of Things: A Survey*

Explainable AI (XAI) offers several advantages in the realm of AI comprehension and ethical considerations. XAI provides a clearer understanding of AI outputs, particularly in deciphering complex decisions made by sophisticated AI models like deep neural networks. Additionally, by addressing the 'AI black box,' XAI contributes significantly to ethical concerns and regulatory compliance, ensuring responsible and accountable AI usage. Moreover, systematic reviews in XAI literature guide future research endeavors, advancing the field's knowledge and applicability. Despite its strengths, XAI encounters limitations that hinder its comprehensive deployment. A significant challenge lies in the limited understanding of how XAI effectively addresses the black-box problem within AI models. Moreover, the identified thematic debates in XAI discussions often lack clarity, impeding a cohesive understanding of XAI's role in improving explainability. XAI research primarily revolves around leveraging sophisticated AI technologies, especially deep neural networks, to interpret and visualize outputs. Additionally, the field emphasizes the need for sociotechnical perspectives, stressing the importance of stakeholder approaches and holistic viewpoints in addressing the explainability of AI systems. while XAI exhibits strengths in enhancing understanding and addressing ethical concerns, challenges persist in achieving a comprehensive understanding and clarity in XAI debates. Research primarily focuses on complex AI technologies and emphasizes the necessity for sociotechnical viewpoints to bolster a comprehensive analysis. Undertaking empirical studies becomes crucial to assess XAI's real-world effectiveness in meeting stakeholder needs and societal expectations.

### *E. Reviewing the Need for Explainable Artificial Intelligence (XAI)*

The AI Explainability 360 toolkit stands out for its comprehensive array of advantages. It amalgamates eight cutting-edge explanation algorithms, tailored to suit the diverse needs of users spanning affected citizens, regulators, experts, and developers. What sets it apart is the introduction of a structured taxonomy, simplifying the navigation of various explanation methods. This taxonomy not only aids practitioners and developers but also reveals gaps, informing crucial design choices for further advancements. Additionally, enhancements in usability for specific algorithms, combined with resourceful tutorials, notebooks, and demonstrations, ensure accessibility for both seasoned practitioners and newcomers to the domain. However, despite these strengths, certain limitations hamper the toolkit's potential. Notably, there are gaps in the coverage of certain explanation categories, particularly in visualizing intermediate nodes/layers and evaluating input feature effects. Furthermore, its potential framework dependence poses a challenge, potentially requiring separate implementations for different deep learning frameworks, restricting universal adoption. Technologically, the toolkit relies on an assortment of advanced methods, leveraging state-of-the-art algorithms to visualize intricate aspects of deep neural networks and assess the effects of input features. Its extensible programming interface fosters flexibility by seamlessly integrating new explanation methods. Educational resources, including tutorials, notebooks, and demonstrations, enrich the toolkit's accessibility, making it user-friendly across diverse audience while the AI Explainability 360 toolkit presents a rich assortment of diverse explanation methods and a structured taxonomy, it grapples with limitations related to method coverage and potential framework dependence. Nonetheless, its technological backbone, featuring advanced algorithms and a flexible interface, supplemented by educational resources, significantly augments understanding and accessibility in the realm of explainable AI.

### *F . One Explanation Does Not Fit All: A Toolkit and Taxonomy of AI Explainability Techniques*

LIME introduces a unique explanation technique that enhances trust in machine learning models by providing interpretable and faithful explanations for predictions. Its flexibility allows it to explain various models across different data types, like text and image classifications, catering to diverse domains and model types. Users can benefit from its explanations in tasks related to trust, making decisions between models, assessing trustworthiness, improving unreliable models, and gaining insights into predictions, proving useful for both expert and non-expert users. While LIME is effective in providing explanations, its primary limitation lies in its focus on sparse linear models for explanations. It may benefit from exploring and comparing other explanation families, like decision trees, to provide a more comprehensive understanding of alternative explanation methods. Additionally, addressing image-based explanations, particularly the "pick" step for images, remains a challenge that requires further attention for refining and enhancing image-based explanation processes. LIME's approach involves learning interpretable models locally around predictions to ensure understandable and faithful explanations. It utilizes submodular optimization techniques to present representative individual predictions and their explanations in a non-redundant manner. Its effectiveness is validated through extensive experimentation involving simulated scenarios and human subjects across various domains, such as text and image classifications, demonstrating its utility and real-world applicability.

DOI: 10.5281/zenodo.10669844

### G .“Why Should I Trust You?” Explaining the Predictions of Any Classifier

GNNEXPLAINER stands out as a pioneering and versatile solution, offering interpretability for predictions made by various Graph Neural Network (GNN)-based models across diverse graph-based machine learning tasks. It introduces a unique capability to identify compact subgraph structures and key node features crucial to GNN predictions, enabling concise and understandable explanations. This approach empowers users with insights into semantically relevant structures and the influential contributions of node features, thus facilitating a more profound understanding of GNN decision-making processes. Moreover, it delivers consistent explanations for entire classes of instances, aiding in systematic comprehension and the identification of errors or anomalies within predictions. However, GNNEXPLAINER might encounter challenges when dealing with larger-scale graphs due to the complexity arising from incorporating both graph structures and node features for explanation. This potential complexity could affect the accuracy and performance of the explanation generation process, limiting its effectiveness, especially in scenarios involving expansive and intricate graph structures. Technologically, GNNEXPLAINER leverages Graph Neural Networks (GNNs) and their recursive neighborhood-aggregation schemes to identify crucial graph pathways and relevant node feature information transmitted along edges, facilitating comprehensive explanation generation. The approach formulates the explanation process as an optimization task, maximizing mutual information between a GNN's prediction and possible subgraph structures. Its effectiveness is validated through extensive experimentation on various synthetic and real-world graph datasets, including MUTAG and BA-COMMUNITY, showcasing its superior explanation accuracy compared to alternative baseline approaches, with improvements of up to 43.0%.

### H .GNN Explainer: Generating Explanations for Graph Neural Networks

Advantages of Object Recognition in Deep Reinforcement Learning (DRL): The integration of object recognition within DRL models serves as a pivotal step toward establishing system transparency. This integration allows the creation of explicit object saliency maps, providing visual representations of the internal states of the system. These saliency maps play a vital role in enabling users to comprehend the decision-making processes of the system, contributing significantly to building trust. Moreover, this transparency facilitates coherent explanations for the actions and decisions executed by the system. However, the augmentation of object recognition within DRL models introduces potential complexities, especially in managing and navigating the increased intricacies of these systems. As the integration enhances transparency, the challenge lies in ensuring the scalability of the DRL models, particularly in more extensive or intricate environments. Balancing this complexity against scalability might pose difficulties, particularly in larger and more complex tasks or domains. The approach predominantly relies on leveraging Deep Reinforcement Learning (DRL) models known for their successful action control in visual domains such as Atari games. Through explicit incorporation of object recognition processing within these DRL models, the framework gains capabilities to enhance transparency, providing valuable insights into the system's decision-making mechanisms. This involves generating object-based saliency maps, which visually represent the internal states of DRLNs, thereby aiding in both explaining the system's decisions and evaluating its actions.

### I. Transparency and Explanation in Deep Reinforcement Learning Neural Networks

Advantages of Object Recognition in Deep Reinforcement Learning (DRL): The integration of object recognition within DRL models serves as a pivotal step toward establishing system transparency. This integration allows the creation of explicit object saliency maps, providing visual representations of the internal states of the system. These saliency maps play a vital role in enabling users to comprehend the decision-making processes of the system, contributing significantly to building trust. Moreover, this transparency facilitates coherent explanations for the actions and decisions executed by the system. However, the augmentation of object recognition within DRL models introduces potential complexities, especially in managing and navigating the increased intricacies of these systems. As the integration enhances transparency, the challenge lies in ensuring the scalability of the DRL models, particularly in more extensive or intricate environments. Balancing this complexity against scalability might pose difficulties, particularly in larger and more complex tasks or domains. The approach predominantly relies on leveraging Deep Reinforcement Learning (DRL) models known for their successful action control in visual domains such as Atari games. Through explicit incorporation of object recognition processing within these DRL models, the framework gains capabilities to enhance transparency, providing valuable insights into the system's decision-making mechanisms. This involves generating object-based saliency maps, which visually represent the internal states of DRLNs, thereby aiding in both explaining the system's decisions and evaluating its actions.

### J .Towards Explainable NLP: A Generative Explanation Framework for Text Classification.

The GEF framework introduces a groundbreaking approach to NLP, enabling the generation of intricate, human-readable

**DOI: 10.5281/zenodo.10669844**

explanations aligned with classification decisions. It excels in producing fine-grained explanations that enhance transparency, aiding users in comprehending model outputs effectively. Furthermore, GEF's unique capability to simultaneously make decisions and generate concise yet contextually relevant explanations stand out, significantly enhancing the interpretability of NLP models. Although GEF demonstrates adaptability across diverse models, its model-agnostic nature might pose integration challenges with various NLP tasks beyond classification. This adaptability issue highlights the need for seamless integration with other NLP functions, such as summarization or extraction, requiring further exploration for comprehensive applicability. GEF harnesses a tailored Generative Explanation Framework combined with the explainable factor and minimum risk training. This amalgamation empowers GEF to generate more coherent and comprehensive explanations, driving improvements in both model performance and the clarity of explanations.

*TABLE I COMPARISONS*

| TITLE | ADVANTAGES | LIMITATIONS | TECHNOLOGY USED |
|---|---|---|---|
| SoK: Applying Machine Learning in Security - A Survey | Enhanced threat detection capabilities through neural networks and decision trees. Automation reducing manual efforts and minimizing false positives. Adaptability to evolving threats facilitated by ML frameworks with big data technologies | ML's dependency on historical data might hinder adaptation to rapidly evolving threats. Publication bias in academic literature might limit exposure to diverse perspectives. Acquiring substantial labeled data for training ML models poses a challenge, especially in security domains. | Diverse ML algorithms, big data frameworks like Apache Spark and Hadoop, deep learning libraries like TensorFlow and PyTorch. |
| Applications of Artificial Intelligence (AI) to Network Security | Adaptability of AI, especially Machine Learning (ML), to evolving threats. Paradigm shift to reactive real-time detection methods powered by AI and ML. Reduction in human intervention through supervised and unsupervised ML | Challenges in understanding AI's actual capabilities versus perceived benefits. Rush to adopt AI without clear comprehension of implications. Need for enhanced data visualization and integration within existing security frameworks. | ML algorithms, deep learning frameworks, big data processing tools, data visualization libraries, security analytics platforms, cloud computing services. |
| XAI for Cybersecurity: State of the Art, Challenges, Open Issues and Future Directions | Transparency and understandable interpretations offered by XAI. Enhanced comprehension in complex decision-making systems across critical domains (healthcare, military, finance) | Conceptual ambiguity leading to misunderstandings and debates in interpreting AI decisions. Absence of standardized measurement metrics affecting universal understanding. Vulnerabilities to adversarial attacks raising concerns about trust and security. | Machine learning algorithms (supervised, unsupervised, reinforcement learning), data analytics and visualization tools, federated learning in IoT networks. |
| Explainable Artificial Intelligence (XAI) for Internet of Things: A Survey | Transparency and understandable interpretations in AI models. Automation simplifying explanations and bolstering security in IoT applications. Contribution to distributed AI decision-making through federated learning in IoT. | Conceptual ambiguity leading to misunderstandings in AI decisions. Lack of standardized measurement metrics affecting universal understanding. Vulnerabilities to adversarial attacks impacting trust and security. | Specialized methodologies tailored to models, automation to ML tasks, federated learning in IoT networks |
| Reviewing the Need for Explainable Artificial Intelligence (XAI) | Clearer understanding of complex AI models, aiding in ethical considerations. Systematic reviews guiding future research endeavors in XAI. | Limited understanding of how XAI effectively addresses the black-box problem within AI model. Identified thematic debates lacking clarity in XAI discussions. | Complex AI technologies (deep neural networks), emphasis on sociotechnical perspectives |
| One Explanation Does Not Fit All: A Toolkit and Taxonomy of AI Explainability Techniques | Introduction of structured taxonomy simplifying navigation of various explanation methods. Accessibility for practitioners and newcomers through educational resources. | Gaps in the coverage of certain explanation categories, potential framework dependence. Complexities in visualizing intermediate nodes/layers and evaluating input feature effects | State-of-the-art algorithms, flexible interface, educational resources like tutorials, notebooks, demonstrations. |
| Why Should I Trust You? Explaining the Predictions of Any Classifier | Providing interpretable and faithful explanations for machine learning model predictions. Assistance in decision-making and assessing trustworthiness of models. | Focus on sparse linear models for explanations, potential improvements in image-based explanations. Addressing challenges related to explaining images. | Learning interpretable models, submodular optimization techniques, extensive experimentation across domains. |
| Transparency and Explanation in Deep Reinforcement Learning Neural Networks | Establishment of system transparency through explicit object saliency maps. Visual representations aiding in comprehending decision-making processes. Facilitation of coherent explanations for system actions and decisions. | Potential complexities in managing increased intricacies of integrated object recognition within DRL models. Balancing complexity against scalability in larger and more complex environments. | Deep Reinforcement Learning (DRL) models incorporating object recognition processing. |
| Towards Explainable NLP: A Generative Explanation Framework for Text Classification | Generation of intricate, human-readable explanations enhancing model transparency. Fine-grained explanations aiding in understanding and decision-making in NLP models. | Model-agnostic nature posing integration challenges beyond classification tasks. Need for seamless integration with other NLP functions like summarization or extraction. | Generative Explanation Framework combined with explainable factor and minimum risk training. |
| GNN Explainer: Generating Explanations for Graph Neural Networks | Identification of compact subgraph structures and key node features crucial to GNN predictions. Consistent explanations aiding in systematic comprehension in graph-based machine learning tasks. | Challenges with larger-scale graphs, complexity affecting accuracy and performance. Incorporating graph structures and node features in explanation might limit effectiveness. | Graph Neural Networks (GNNs), recursive neighborhood-aggregation schemes, extensive experimentation on synthetic and real-world graph datasets. |

## *III . CONCLUSION*

The convergence of Explainable AI (XAI) with security domains and technological advancements signifies a promising trajectory towards transparency, robustness, and comprehension in AI systems. As this review illuminates the spectrum of challenges, potentials, and technological foundations, it becomes evident that fostering explainability in AI models is pivotal for trust-building, innovation, and responsible deployment. The intricate interplay between XAI and security, coupled with technological intricacies, underlines the necessity for continuous research, cohesive standards, and adaptable frameworks to navigate the evolving landscape of AI-driven solutions. This holistic understanding underscores the pivotal role of XAI in shaping the future of secure, transparent, and reliable AI ecosystems across diverse domains.

**DOI: 10.5281/zenodo.10669844**

## REFERENCES

[1]. H. Jiang, J. Nagra, and P. Ahammad, ‘‘SoK: Applying machine learning 2183 in security—A survey,’’ Nov. 2016, arXiv:1611.03186.

[2]. G. Srivastava, R. H. Jhaveri, S. Bhattacharya, S. Pandya, 2235 P. K. R. Maddikunta, G. Yenduri, J. G. Hall, M. Alazab, and 2236 T. R. Gadekallu, ‘‘XAI for cybersecurity: State of the art, challenges, 2237 open issues and future directions,’’ 2022, arXiv:2206.03585

[3]. J. Gerlings, A. Shollo, and I. Constantiou, ‘‘Reviewing the need for 2293 explainable artificial intelligence (XAI),’’ 2020, arXiv:2012.01007.

[4]. Z. A. E. Houda, B. Brik, and L. Khoukhi, ‘‘‘Why should i trust your IDS?’: 2870 An explainable deep learning framework for intrusion detection systems 2871 in Internet of Things networks,’’ IEEE Open J. Commun. Soc., vol. 3, 2872 pp. 1164–1176, 2022, doi: 10.1109/OJCOMS.2022.3188750.

[5]. R. Ying, D. Bourgeois, J. You, M. Zitnik, and J. Leskovec, ‘‘GNNEx2330 plainer: Generating explanations for graph neural networks,’’ 2019, 2331 arXiv:1903.03894.

[6]. R. Iyer, Y. Li, H. Li, M. Lewis, R. Sundar, and K. Sycara, ‘‘Transparency 2338 and explanation in deep reinforcement learning neural networks,’’ 2018, 2339 arXiv:1809.06061

[7]. H. Liu, Q. Yin, and W. Yang Wang, ‘‘Towards explainable NLP: 2352 A generative explanation framework for text classification,’’ 2018, 2353 arXiv:1811.00196.

[8]. A. P. Veiga, ‘‘Applications of artificial intelligence to network security,’’ 2191 2018, arXiv:1803.09992

[9]. M. Danilevsky, K. Qian, R. Aharonov, Y. Katsis, B. Kawas, and P. Sen, 2355 ‘‘A survey of the state of explainable AI for natural language processing,’’ 2356 2020, arXiv:2010.00711.

[10]. V. Arya, R. K. E. Bellamy, P.-Y. Chen, A. Dhurandhar, M. Hind, 2318 S. C. Hoffman, S. Houde, Q. Vera Liao, R. Luss, A. Mojsilović, 2319 S. Mourad, P. Pedemonte, R. Raghavendra, J. Richards, P. Sattigeri, 2320 K. Shanmugam, M. Singh, K. R. Varshney, D. Wei, and Y. Zhang, ‘‘One 2321 explanation does not fit all: A toolkit and taxonomy of AI explainability 2322 techniques,’’ 2019, arXiv:1909.03012