SAND2021-6069C

Considerations for Deploying xAI Tools in the Wild

Lessons Learned from xAI Deployment in a Cybersecurity Operations Setting

Megan Nyre-Yu
Sandia National Laboratories
Albuquerque, NM, USA
mnyreyu@sandia.gov

Blake Moss
Sandia National Laboratories
Albuquerque, NM, USA
bmoss@sandia.gov

Michael R. Smith
Sandia National Laboratories
Albuquerque, NM, USA
msmith4@sandia.gov

Elizabeth S. Morris
Sandia National Laboratories
Albuquerque, NM, USA
esmorri@sandia.gov

Charles Smutz
Sandia National Laboratories
Albuquerque, NM, USA
csmutz@sandia.gov

ABSTRACT

The rapid growth of artificial intelligence (AI) applications in human-interactive settings is spurring development of explainable AI (xAI) techniques. Such efforts require not only robust and reliable models, but also relevant and understandable explanations for end users to successfully assist in achieving user goals, reducing bias, and improving trust. Cybersecurity operations settings represent one such context in which automation is vital for maintaining cyber defenses. AI models and xAI techniques were developed to aid analysts in identifying events and making decisions about each of the flagged events (e.g. benign, malicious). We found that xAI tools, while increasing trust, were not utilized heavily nor did they improve analyst decision accuracy. In deploying the xAI tools, critical lessons were learned that impact the utility and adoptability of the technology, including consideration of end users, their workflows, their environments, and their propensity to trust xAI outputs in their respective roles.

CCS CONCEPTS

• **Human-centered computing** ~Human computer interaction (HCI) ~Empirical studies in HCI • **Computing methodologies** ~Artificial intelligence

KEYWORDS

cybersecurity; explainable AI; security; incident response; deployment; user study

ACM Reference format:

Megan Nyre-Yu, Elizabeth S. Morris, Charles Smutz, Blake Moss, and Michael R. Smith. 2021. Considerations for Deploying xAI Tools in the Wild: Lessons Learned from xAI Deployment in a Cybersecurity Operations Setting. In *Measures and Best Practices for Responsible AI at KDD 2021 (Responsible AI @KDD 2021). KDD 2021, Singapore/Virtual, 4 pages.* https://doi.org/xxx

1 Introduction

Rapid improvements in artificial intelligence (AI) techniques have resulted in significant increases in their usage in a diverse and expanding set of applications. While original successes were in domains with fairly low consequences such as product and movie recommendations, AI algorithms are being used in increasingly higher-consequence applications such as malware detection [1], autonomous vehicles [2], and medical diagnoses [3]. Wide-spread use is limited, however, as there is a recognized need to trust and understand the AI models before they are deployed and integrated into larger systems. In response, several explainable AI (xAI) techniques have emerged [4] to build trust and ensure that the models are fair.

Applying AI models in cybersecurity operations settings is a new and growing area, with strong emphasis on overcoming inefficiencies and uncertainties and improving overall incident response performance [5]. Cyber-attacks result in loss of monetary resources and/or system resource availability. These attacks are increasing in volume and sophistication. AI methods offer improvement to the defense of cyber infrastructure, resulting in the preservation of significant resources. AI has been investigated in several cyber domains including malware detection [6, 7] and malicious PDF detection [8]. xAI has been examined systematically using deep learning methods in cyber defense [9, 10], but independent of the cybersecurity analyst.

We examined the use-case of AI models with explanations for identifying malware in a computer network defense setting. Attacks that occur on enterprise networks are of such a large scale that automated techniques are needed to help manage the attacks. Given the high impact of false negatives, cybersecurity analysts are highly skeptical of automated tools. To increase the productivity of the cybersecurity analysts, the AI model not only needs to be robust and reliable, but also the cybersecurity analyst needs to trust the model to make effective use of its output. However, AI and xAI

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology & Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

methods are often deployed without evaluating how they affect the overt decision process.

This paper details a case study examining the usefulness of xAI techniques integrated into the workflow of cybersecurity analysts in a live setting. In this setting, the cybersecurity analysts need to not only identify malicious artifacts, but also provide reasons *why* they are malicious. Hence, the goal of providing xAI methods is two-fold: to help scale with the increasing number of malicious attacks *and* to point to why the artifact is malicious as part of a cybersecurity analyst's workflow.

2 Motivation and Research Goals

To assess human decision making when presented with model explanations, a user study was conducted with a broader population beyond cybersecurity analysts [11]. The study revealed that when making a decision about a potentially malicious stimulus participants often agreed with the xAI outputs, indicating high inherent trust in the model. We also found that the number of features presented was not a significant factor in the decision to agree with the model's recommendation or not.

The next step was to understand the use of xAI in a cybersecurity context with real analysts. To evaluate the effectiveness of the model and explanations, we planned to collect objective and subjective measures from actual end users in a live security setting. We planned to compare decision behaviors of analysts before and after the xAI deployment to determine if, and how much, cybersecurity analysts trusted and/or used the outputs.

When deploying these techniques in the real world, there were many decision points and limiting or confounding factors that had to be considered to conduct an evaluation of the new tool. This paper describes our evaluation and deployment of the xAI tool, as well as the lessons learned along the way about how cybersecurity analysts in general interact with xAI in real time. This paper does not focus on the visualization methods and design by which the xAI tool would display information to the human user. Rather, we present findings related to practical deployment of the tool. We also provide a list of considerations for AI developers and explanation designers that will help guide decisions during this process. Here we used TreeSHAP [12], but our results are not specific to TreeSHAP and any xAI tool that provides feature importance for a prediction could be used.

Research Question: What practical considerations should be taken into account when developing and deploying AI and xAI tools in high-consequence, live settings?

3 Methods

3.1 Evaluating Effectiveness of AI Tools

Cybersecurity analysts working in real-world incident response teams must make fast triage decisions using multiple pieces of information often including AI model outputs. In this use case, cybersecurity analysts triage multiple alerts to determine if flagged activity is actually malicious. Our goal was to evaluate the efficiency and effectiveness of a single AI model output in the context of incident handling before and after an xAI tool was introduced. We collected (1) instrumented data throughout each of two time periods: pre-xAI tool and post-xAI tool deployment, and (2) survey data from analysts after deploying the xAI tool. A visualization of the current xAI tool is shown in Figure 1.

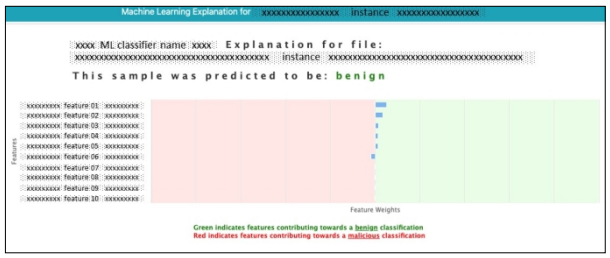

Figure 1. Obscured representation of the xAI tool output when expanded by analyst.

3.1.1. Instrumented data collection. The goal of the user study is to better understand the process used by cybersecurity analysts when promoting an alert and whether the analyst complied with AI model output. Data collection was programmed on the backend of existing cybersecurity tools to prevent the interruption of analyst workflow. Instrumented data collection included when/if an alert was promoted, model output, how the cybersecurity analyst interacted with the alert, and other activities performed on each alert. For the post time period, data indicating whether an analyst opened the xAI tool was also collected.

3.1.2. Subjective trust and explainability of usefulness. To understand the trust level of and satisfaction with explanations from xAI, end user perceptions were measured in a survey via two scales: the Trust Scale Recommended for xAI and the xAI Explanation Satisfaction Scale [13]. The Trust Scale measures whether end users are confident in the xAI tool, and whether the xAI tool is predictable, reliable, efficient, and believable. The Explanation Satisfaction Scale captures end users' judgments about the xAI tool. The cybersecurity analysts were invited to complete an online, 16-item questionnaire including these two scales after at least one week of working with the xAI tool.

3.2 AI Tools in a Live Security Setting

We identified some important attributes about the operational cybersecurity environment we studied to provide some context. First, there is a high cost of undetected malware. Intrusion detection systems are tuned to be sensitive because the cost of undetected malware can be extremely high. Second, intrusion detection systems include multiple, sometimes partially overlapping, alerting criteria. Third, within this context there is a bias towards hard cases; easily detectable malware is automatically mitigated with existing tools and, therefore, not triaged. Samples triaged by analysts are harder to classify and often involve contradictory predictions from competing (and highly accurate) mechanisms. Fourth, to automatically process files with AI algorithms, there is a semantic gap between real-world interpretation and low-level feature space for learning-based intrusion detection systems [14]. In other words,

the interpretation of feature space is not self-apparent (such as is the case with some image classification problems) [15]. Finally, in the team we studied, all incident responders operate on a level organizational structure (versus tier-based structures seen in many security operations). Notably, the individuals who triage alerts are largely disjointed from individuals who maintain the AI models.

It was very challenging to collect data in a scientific way to assess the usefulness and efficacy of using xAI. This is a known challenge in cybersecurity operations settings [16], and we adopted knowledge already learned when constructing our hypotheses and initial research questions about the tool. However, as we devised the plan for collecting data towards answering those research questions, we discovered additional factors that refuted initial assumptions about how the tool would be used. Factors included:

3.2.1 Decision task and alternate decisions support paths. The analysts use the classification output from the AI model with other alert data to make a decision about an event; they may not regularly question the classification output. To mitigate this, we captured data from before/after the tool was deployed to see if including explanations changes analyst behaviors. We seek to capture measurements such as if the section that contains the explanations is expanded and average response time. We also made the visual presentation of explainability more palatable compared to previous versions, which did not organize or present explanations in ways that could be quickly utilized during the decision-making process.

3.2.2 Workflow. Much of the information cybersecurity analysts use to make a triage decision exists in a central incident handling tool, with little navigation required within the dashboard to find decision-critical information. This is a standard workflow in cybersecurity operations settings, and thus perhaps less critical for our own studies. However, user workflow should be considered prior to deployment of xAI techniques in some fashion to understand potential friction points for adoptability.

3.2.3 Tool separation/location. The xAI tool exists outside the main dashboard where analyst conclusions are registered; it is located in a supporting software program which requires pivoting to engage with it. While this program is routinely accessed by analysts, the addition of the tool was not immediately obvious. To mitigate this, we (1) hosted training with the analysts so they would be able to locate the xAI tool, and (2) created an interface feature (Figure 1) to increase salience of the new xAI tool.

3.2.4 Number of end users and their roles. In our scenario, there is an assigned primary incident responder per week causing turnover and rotation within the group of users whose roles differ regarding decision-making about an event. To mitigate this, we include all users who might interact with the explainability tool, not just the incident responders who are primarily responsible for incidents in a given week.

4 Results

4.1 Data Findings

We collected instrumented data without interrupting the incident handling context with and without xAI tools. We then measured user trust and perception of usefulness of the xAI tool.

4.1.1. Instrumented data findings. As described in 3.1, data were collected continuously over the course of several months; we monitored this data stream to capture a pre-deployment baseline of existing tool use and post-deployment data to ensure the xAI tool was working properly. Surprisingly, we discovered that users were not interacting much with the existing tool or the new tool. Why were these analysts not opening the explanations? A shift in thinking allowed us to appreciate the key finding in the pre-deployment data: the targeted analysts did not use explanations in their daily workflows. Moreover, the placement of a new tool in an inconspicuous location will further decrease the likelihood that users engage with the tool and relying on training to overcome that limitation is an insufficient strategy.

4.1.2. Subjective trust and usefulness data findings. As of the writing of this paper in May 2021, only one cybersecurity analyst completed the two scales included in the questionnaire. Analysts were provided multiple opportunities and reminders to complete the 16-item questionnaire. Because the experimental environment was a live setting, there is a greater risk of non-compliance; that is, a questionnaire is viewed as an interruption to a cybersecurity analyst's workflow. Thus, a key component in researching xAI tools for deployment in the wild is to make the experience as non-cumbersome as possible. While most acknowledge the importance of AI in keeping up with cyber threats, in practice the additional overhead in an overwhelming domain produced little utility.

4.2 Deployment Challenges

Despite our efforts to understand analyst interaction with the system using unobtrusive data collection, we faced several challenges in deploying the tool in a live setting. Due to the chosen location of the tool, we expected some level of low engagement. To mitigate this risk of low familiarity with the tool's existence, we conducted a single-day training, which covered an overview on the tool's user interface, as well as a tutorial on its operational use. However, not all analysts were able to attend the training, and some analysts identified this as the reason they did not use the tool.

The explanation capability was added to existing intrusion detection systems with the assumption that understanding model rationale would help with triage tasks. One of the core insights gained is that *this is a false premise*. Taking the time to understand the rationale of one of many possible, and often contradictory, detection mechanisms is not necessarily the most efficient path for triage. This is especially true when analysts have other sources of information available to them that are more easily consumable, including the observation itself (e.g., the file that might be malware) and data views that have been developed based on analyst feedback. To some degree, performing the same manual analytic steps on samples regardless of alert source might ensure consistency and help prevent analytical bias.

The xAI tool targeted analysts based on the hypothesis that improved understanding of the model's decisions would increase analyst confidence and improve overall performance. Due to widespread skepticism amongst security analysts, this hypothesis made sense: provide more data such that their skepticism is satisfied. However, we believe that injecting a new xAI tool over

existing models that analysts already trusted impact our ability to detect gain in confidence and reduction of skepticism.

We were also prepared to face challenges related to the environment in which the xAI tool was deployed. Though we found this to be not as relevant for our use case, the context of the xAI deployment may impact its usefulness and adoptability. Security incident responders are known to experience high load of alerts, and are subject to different kinds of cognitive biases when interacting with intrusion detection systems [17]. These settings have a history of high turnover and burnout [18, 19], and judgments about alerts are often made with pressure from long queue of alerts or time expected to make a decision [20]. More relevant to our use case was that the actual workflow of incident response analysts did not include validation of detection mechanism outputs (xAI or not), and rationalizing those outputs is not an efficient path.

5 Discussion

Our own efforts resulted in some lessons learned for deploying AI models "in the wild", which might be useful for others developing xAI for use in real-world settings. The study originally aimed to conduct more controlled field experiments, which quickly evolved into tool improvement. Over the period of about 6 months, we were forced to modify our original study design to the extent that we developed new research questions and pursued entirely new studies. For some researchers, these changes represent some level of risk, which we believe can be mitigated by learning from studies like ours and considering certain design and deployment elements before commencing data collection.

We learned that our hypothesized user base seemingly trusts the output of the AI model to the extent that they do not explore provided novel explainability tools, similar to previous conclusions on non-expert users [11]. Further investigation revealed that incident responders, or the people who are *making decisions* from the AI outputs (and a suite of other tools), are not interested in validating the model. However, this realization led us to consider two new questions: (1) who would validate the model outputs, and thus potentially benefit from xAI tools, and (2) how can the incident responders still contribute to the quality of AI explanations?

We found that AI model maintainers, or the experts tasked with training AI models and monitoring their performance, are more invested in *verifying model outputs* for the purpose of improving model accuracy. Accordingly, we have pivoted our efforts to understand that user base better. Future research also includes exploring the second question of increasing input from incident responders into AI tools without interrupting normal workflow.

We also learned that the context in which the xAI tool was deployed dictated how much it might actually be used and for what purposes. Though most of this lesson was learned through literature review, we still found that some aspects of context impacted how we deployed the xAI tool. For instance, we considered how the presentation of information in the explanations could be improved such that the outputs were meaningful to the target users. Additional contextual factors, such as time pressure, task volume, and consequences of trusting xAI tools, were found to be less critical for our use case but should be considered for researchers and developers planning to deploy such tools in real environments.

Based on the above lessons learned, we offer considerations for developing and deploying xAI tools in live contexts. The following questions can help guide decisions and mitigate risks during various stages of technology transfer.

Table 1. Practical considerations for xAI deployment

Practical Consideration	Supporting Questions
Who are your end users?	Who uses the model outputs, and in what way?
	How does the xAI tool help them accomplish their goals?
	With respect to explainability, who critically questions how the model works (within their normal workflow)?
What is the context in which the model is deployed?	Do environmental pressures counteract the availability of the model?
	Are the features, feature names, and visual representations of explainability relevant and meaningful in this context?
What is the relative risk of the model being wrong?	How does the risk of model inaccuracy impact the end user?
	What are the consequences of trusting the model?
What is the risk of the explanation being unclear or incorrect?	How does an unclear explanation impact the end user?
	What are the consequences of presenting a poor or incorrect explanation?

6 Conclusions

While conducting a study aimed to understand if cybersecurity analysts would benefit from an xAI tool in a real-world setting, we learned that a team of cybersecurity analysts seemingly trust the output of the AI model and do not explore the provided explanations. Rather, other existing tools are used to validate the output of AI models. In this context, the use of the output from the AI classification model was embedded in cybersecurity analysts' main workflow, while the use of the new xAI tool was not.

We identified considerations that researchers and developers can integrate into current processes to design and target better xAI tools for more successful technology transfer. Additionally, examining real-time, nonintrusive data from instrumented backend data collection is a great means to understand if and how end users are using an xAI tool. Ultimately, considering the end users and their context early in the process reduces risks and impact of unidentified challenges.

ACKNOWLEDGMENTS

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology & Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525. This paper describes objective technical results and analysis. Any subjective views or opinions that might be expressed in the paper do not necessarily represent the views of the U.S. Department of Energy or the United States Government.

REFERENCES

[1] Asaf Shabtai, Uri Kanonov, Yuval Elovici, Chana Glezer, and Yael Weiss. 2012. "Andromaly": a behavioral malware detection framework for android devices," *Journal of Intelligent Information Systems,* vol. 38, no. 1, 161-190.

[2] Hubert Etienne. 2020. "When AI Ethics Goes Astray: A Case Study of Autonomous Vehicles," *Social Science Computer Review*. February 2020.

[3] Bradley J. Erickson, Panagiotis Korfiatis, Zeynettin Akkus, and Timothy L. Kline. 2017. "Machine Learning for Medical Imaging," *RadioGraphics,* vol. 37, no. 2, 505-515.

[4] Alejandro B. Arrieta *et al.* 2020. "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI," *Information Fusion,* vol. 58, 82-115.

[5] Constantin Nilă, Ioana Apostol, and Victor Patriciu. 2020. Machine learning approach to quick incident response. In *2020 13th International Conference on Communications (COMM)*, June 18-20, 2020, Bucharest, 291-296.

[6] Hyrum S. Anderson and Phil Roth. 2018. "Ember: an open dataset for training static pe malware machine learning models," *arXiv preprint arXiv:1804.04637*.

[7] Edward Raff, John Barker, Jared Sylvester, Robert Brandon, Bryan Catanzaro, and Charles Nicholas. 2017. "Malware detection by eating a whole exe," *arXiv preprint arXiv:1710.09435*.

[8] Charles Smutz and Angelos Stavrou. 2012. Malicious PDF detection using metadata and structural features. In *Proceedings of the 28th Annual Computer Security Applications Conference*, December 3-4, 2012, Orlando, Florida, 239-248.

[9] Aditya Kuppa and Nhien-An Le-Khac. 2020. Black Box Attacks on Explainable Artificial Intelligence (XAI) methods in Cyber Security. In *2020 International Joint Conference on Neural Networks (IJCNN)*: IEEE, 1-8.

[10] Alexander Warnecke, Daniel Arp, Christian Wressnegger, and Konrad Rieck. 2020. Evaluating explanation methods for deep learning in security. In *2020 IEEE European Symposium on Security and Privacy (EuroS&P)*, 2020: IEEE, 158-174.

[11] Mallory C. Stites, Megan Nyre-Yu, Blake Moss, Charles Smutz, and Michael R. Smith. 2021. Sage advice? The impacts of explanations for machine learning models on human decision-making in spam detection. In *Proceedings of the HCI International Conference, HCII*, Washington, DC, USA.

[12] Lundberg, S.M., Erion, G., Chen, H. et al. From local explanations to global understanding with explainable AI for trees. *Nat Mach Intell*, **2**, 56–67 (2020).

[13] Robert R. Hoffman, Gary Klein, and Shane T. Mueller. 2018. Los Angeles, CA. Explaining Explanation For " Explainable AI ".

[14] Michael R. Smith, Nicholas T. Johnson, Joe B. Ingram, Armida J. Carbajal, Bridget I. Haus, Eva Domschot, Ramyaa Ramyaa, Christopher C. Lamb, Stephen J. Verzi, and W. Philip Kegelmeyer. 2020. Mind the Gap: On Bridging the Semantic Gap between Machine Learning and Malware Analysis. In *Proceedings of the 13th ACM Workshop on Artificial Intelligence and Security (AISec'20)*. Association for Computing Machinery, New York, NY, USA, 49–60.

[15] Robin Sommer and Vern Paxson. 2010. "Outside the Closed World: On Using Machine Learning for Network Intrusion Detection. In *2010 IEEE Symposium on Security and Privacy*, May 16-19, 2010, 305-316.

[16] Robert R. Hoffman. 2019. "The Concept of a "Campaign of Experimentation" for Cyber Operations," *The Cyber Defense Review,* vol. 4, no. 1, pp. 75-84.

[17] Antione Lemay and Sylvain Leblanc. 2018. Cognitive Biases in Cyber Decision-Making. In *Proceedings of the 13th International Conference on Cyber Warfare and Security,* March 8-9, 2018, Washington, D.C., 395-401.

[18] JP Bourget. 2017. Addressing Analyst Fatigue in the SOC. Arlington, VA: https://www.brighttalk.com/webcast/288/224207.

[19] Christina Richmond and Pete Lindstrom. 2015. "IDC Security Survey: As the Job Churns".

[20] Christopher Petersen and Robert Lentz. 2015. Surfacing Critical Cyber Threats Through Security Intelligence: A Reference Model for IT Security Practitioners. https://www.ciosummits.com/A4_LR_SIMM_CISO_Whitepaper.pdf