Explainable AI in Cybersecurity Operations: Lessons Learned from xAI Tool Deployment

Megan Nyre-Yu
and Elizabeth Morris
Statistics and Human Systems
Sandia National Laboratories
{mnyreyu, esmorri}@sandia.gov

Michael R. Smith
All-Source Analytics
Sandia National Laboratories
msmith4@sandia.gov

Blake Moss
and Charles Smutz
Cyber Security Technologies
Sandia National Laboratories
{bmoss, csmutz}@sandia.gov

Abstract—**Technological advances relating to artificial intelligence (AI) and explainable AI (xAI) techniques are at a stage of development that requires better understanding of operational context. AI tools are primarily viewed as black boxes and some hesitation exists in employing them due to lack of trust and transparency. xAI technologies largely aim to overcome these issues to improve operational efficiency and effectiveness of operators, speeding up the process and allowing for more consistent and informed decision making from AI outputs. Such efforts require not only robust and reliable models but also relevant and understandable explanations to end users to successfully assist in achieving user goals, reducing bias, and improving trust in AI models. 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 flagged events (e.g. network attack). We instrumented the tools used for cybersecurity operations to unobtrusively collect data and evaluate the effectiveness of xAI tools. During a pilot study for deployment, we found that xAI tools, while intended to increase trust and improve efficiency, were not utilized heavily, nor did they improve analyst decision accuracy. 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.**

I. 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 medical diagnoses [3]. Widespread use is limited, however, as there is a recognized need to trust and understand the decision processes of AI models before they are deployed and integrated into larger systems. In response, several explainable AI (xAI) techniques have emerged [1] to build trust and ensure that a model is not biased.

Using AI models in cybersecurity operations settings is growing, as it promises a way to manage increasing traffic and cyber attacks. Cyber attacks result in significant loss of monetary resources and/or system resource availability. AI methods offer improvement to defense of cyber infrastructure, running at machine speeds and resulting in preservation of significant resources. AI has been investigated in several cyber domains, including malware detection [12] and malicious PDF detection [16]. xAI has been examined systematically using deep learning methods in cyber defense [19], but independent of the cybersecurity analyst. Our goal was to evaluate how xAI tools affect cyber analysts in their daily workflow.

A. Study scope

We examined the use case of AI models with explanations for identifying malware in a live computer network defense setting with human operators. Given the high impact of false negatives, cybersecurity analysts are highly skeptical of automated tools. To increase the productivity of the cybersecurity analysts, not only does the AI model need 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 methods are often deployed without evaluating how they affect the overt decision process. Moreover, if the hope is to deploy these technologies successfully, technology adoption measures such as usefulness and usability should be evaluated [9].

This paper presents a case study examining the use of xAI techniques integrated into the workflow of cybersecurity analysts. In this setting, the cybersecurity analysts need to both identify malicious artifacts and provide reasons why they are malicious. Hence, the goal in providing xAI methods is twofold: 1) to help analysis scale with the increasing number of malicious attacks and 2) to point to why an artifact is malicious as part of a cybersecurity analyst's workflow. The usefulness of the xAI tool was evaluated through several data collection methods and found to be less useful to cyber analysts than originally hypothesized. While the deployment of the xAI tool was considered a failure within an incident handling task, we identified another population within the cyber context that showed interest in using the tool. We share valuable lessons learned about design and deployment of new tools in cybersecurity contexts at large. Moreover, we share aspects of our data collection methodology, which utilized system-based instrumentation that did not interfere with a cybersecurity analyst's current workflow and could be useful to other usability researchers and automation developers in the cyber domain.

Usable Security and Privacy (USEC) Symposium 2022
28 March 2022, Virtual
ISBN 1-891562-79-7
https://dx.doi.org/10.14722/usec.2022.23014
www.ndss-symposium.org, https://www.usablesecurity.net/USEC/

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.

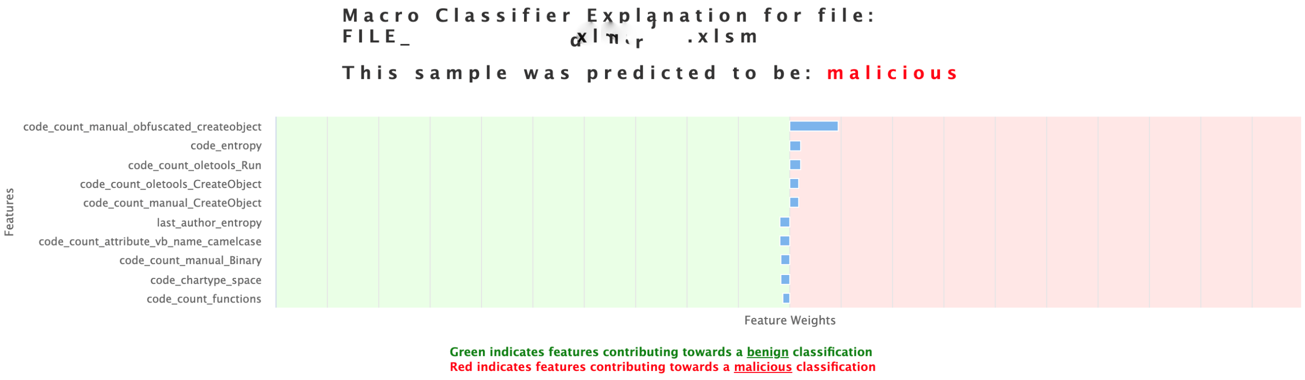

Fig. 1. Example of the explanations from the xAI tool when expanded by analyst. (Note that the file name has been purposely obfuscated).

B. *Motivation and research goals*

To assess human decision making when presented with model explanations, an initial study was conducted with a broader population beyond cybersecurity analysts [18]. The study revealed that when making a decision about a potentially malicious stimulus, participants often agreed with the AI outputs, indicating high inherent trust in the model regardless of the output from the xAI tool. We also found that the number of model features presented was not a significant factor in the decision of whether or not to agree with the model's recommendation.

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 quantify how much the efficiency of cybersecurity analysts increased or decreased when triaging suspicious events.

When deploying these techniques in the real world to conduct an evaluation of the xAI tool, many decision points and confounding factors were considered. This paper describes our evaluation and deployment of the xAI tool, our instrumentation for collecting data, and lessons learned about how cybersecurity analysts 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 user. Rather, we present findings related to practical deployment of the tool. We also provide a list of considerations for AI developers and xAI designers that will help guide decisions during this process. Here we used TreeSHAP [8], but any xAI tool that provides feature importance for a prediction could be used.

Research Question: What practical considerations are necessary when developing and deploying AI and xAI tools in live, high-consequence settings?

II. Methods

A. *Evaluating effectiveness of AI and xAI tools*

Cybersecurity analysts working in real-world incident response teams must make quick triage decisions using multiple pieces of information, often including the outputs from AI models. In this use case, cybersecurity analysts triage multiple alerts to determine if flagged suspicious activity is actually malicious. Our goal was to evaluate the effectiveness and efficiency of a single AI model output in the context of incident handling before and after an xAI tool was introduced. We collected (1) instrumented (log) 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 screenshot showing sample output of the current AI output with explanations from an xAI tool is presented in Figure 1.

1) User population and task: The purpose of our study was to better understand how new technologies, such as AI and specifically xAI, can be developed and deployed in real cyber settings. Thus, our study was not considered human subjects research (HSR) by the institutional review board (IRB) because our study implemented a tool within an existing process for the purpose of process improvement. Accordingly, our results highlight valuable lessons learned in the development and deployment process, not performance-related data of the human analysts using the AI and xAI tools. The group of cyber analysts was comprised of eleven ($n = 11$) individuals who rotated responsibilities for incident handling. These analysts were part of an existing, established team operating within a large company with established enterprise security protocols and practices. All analysts used common tools and resources to conduct incident handling tasks. While this sample may seem small, it is considered normal for this context; there is a wide range of "normal" team sizes in security operations [14].

Our study focused on a common incident handling task: determining whether an automated alert is a threat that needs to be addressed or was adequately addressed by automated systems. This is a decision making task to determine if an "alert" should be promoted for further action, with the underpinning decision of whether the alert is malicious or benign. This task requires context switching [4] across a variety of tools and systems [14], [10]. Our study injected a new xAI tool into existing software used by analysts to observe how it affected their decisions and behaviors while making this decision.

2) The xAI tool and instrumented data collection: The xAI tool, pictured in Figure 1, presents a ranked list of the most influential features for the model's prediction for a given file

TABLE I. SYSTEM INSTRUMENTATION FOR CAPTURING RELEVANT USER INTERACTIONS.

Instrumented Variable	Type	Method	Location
Interaction with specific interface feature – Hover-over	Proactive	JavaScript addition to source code	Object Scanning application
Interaction with specific interface feature – Mouse-click	Proactive	JavaScript addition to source code	Object Scanning application
Time an analyst viewed	Retroactive	Pre-existing	Incident Tracking System
Analyst ID (alias)	Retroactive	Usernames are always attached to audit records in both the object scanning and incident tracking platforms. After the logs were collected, these usernames were substituted for a cryptographic hash of the username with a concatenated random string (alias)	N/A
If promoted to "event" status	Retroactive	Pre-existing	Incident Tracking System

(PDF or macro). The intent of this visualization was to help focus an analyst's efforts on the most important attributes of the file to make triaging and validating (determining ground truth) observations more efficient. The items listed along the y-axis are model features that are relevant for predicting the outcome, while the x-axis shows strength of that feature for the artifact. Greater magnitude to the right (shown as red) indicates stronger alignment with known malicious features, while magnitude to the left (shown as green) indicates stronger alignment with benign features. This format is commonly used in xAI applications. For the artifact shown in the figure, the classifier outcome of "malicious" is presented to the analyst in the object scanning tool, which can be further explored through this xAI tool to understand why the model predicted this artifact as malicious.

Data collection was programmed on the back-end of existing cybersecurity tools primarily to *prevent workflow interruption.* Instrumentation, or data collecting mechanisms within the system and/or software, is an unobtrusive method that is important to consider in cybersecurity contexts for several reasons. First, the task being conducted is often measured in terms of time; data collection methods that minimize or eliminate interruption (e.g., surveys, verbal protocols, etc.) should be considered to help minimize impact on performance. Second, common HSR concerns about identifiability can be easily avoided through system-instrumented data collection. Using this approach helps qualify studies for exempt status, greatly reducing the time (and resulting risk) for IRB approval. The authors did seek IRB approval, but this study was considered non-HSR due to the fact that it was a usability study on improving an existing tool. Third, as a very practical concern, many security contexts are wary of outsiders, and building trust and rapport with analysts is challenging. It is also difficult to get direct access to these individuals to pursue traditional approaches for studying technology adoption. System-instrumented data collection reduces or removes the need for researchers and developers to directly interact with analysts when testing a tool. Lastly, a benefit of system instrumentation is the simultaneous accomplishment of building a data collection network that could help current or future automation development, fueling automated systems with live data from users to understand their actions and possibly infer intent.

Accordingly, we provide some details regarding our methods and decisions for instrumented data collection for others who would like to use this method in cybersecurity contexts. Our instrumentation included model output, when/if an alert was promoted, how a cybersecurity analyst interacted with the alert, and other activities performed on each alert. As was mentioned previously, data were collected "pre" deployment of the xAI tool and "post" deployment. For the "post" time period, data indicating whether an analyst opened the xAI tool was also collected.

Incident response teams use two important types of systems: scanning platforms and incident tracking systems. Scanning platforms can take a variety of forms but almost always involve automation to extract relevant information from digital artifacts (files, packet captures, binaries, etc.). The xAI tool was implemented in a scanning platform to succinctly provide analysts with information regarding why a specific artifact was classified as malicious or benign. The tool reports which features were important for determining that classification. Additionally, incident tracking systems are used in operations to help incident response teams record and coordinate investigations and responses to security incidents. These systems present digital artifacts that meet various alerting criteria to an analyst. The analyst can then investigate and add notes and other relevant information during the investigation and remediation phases. If initial investigations yield substantial findings, incidents are escalated or promoted in the tracking system (i.e. "event") for further, deeper analysis. These incident tracking systems are analogous to general service desk/ticket management platforms ubiquitous in the information technology domain.

We examined different approaches that might be helpful for experiment designers and application designers to collect relevant data. It is important to be cognizant of granularity of user interactions to be collected for answering specific research questions. Data collection approaches that collect copious amounts of data without specific research questions can overwhelm a data analyst with too much data and result in few, if any, useful insights.

Our processes for collecting user interaction metrics for the xAI tool involved both *proactive* and *retroactive* instrumentation to the scanning platform and incident tracking system used by the cybersecurity team. A summary of the instrumentation used for this case study is shown in Table I. None of the modifications changed the workflow for the analysts.

Proactive instrumentation captures specific data not included in the application's existing audit functionality and was implemented before the experiment. Audit functionality was proactively added to front-end and back-end code to record mouse-click and cursor-hover user interactions with the xAI tool and adjacent user interface features (e.g. accordion menus, displayed information, tooltips, etc). This was done by modifying the front-end (JavaScript) and back-end (Python) source code of the in-house developed object scanning application.

Retroactive instrumentation leverages data that was already being collected automatically by analyst tools. To gather a more complete view of the incident response life cycle [14] in relation to the xAI tool, retroactive auditing was implemented in the incident tracking platform using custom python scripts that pull existing audit records and calculate base statistics. This auditing filtered for alerts generated by digital artifacts that could be tied to xAI outputs. It then recorded various life cycle events in the incident tracking system such as: when/which analysts viewed the incident and if the incident was elevated to a heightened status. We relied on existing auditing functionality in the incident tracking application to capture this information and utilized batch processing (Python scripts) to extract the relevant audit records from the application's database.

Finally, refined audit records from both applications were grouped by digital artifact. This provided an overarching timeline of analyst interactions with the xAI tool, object scanning, and incident tracking platforms with respect to each digital artifact. This was also implemented using custom Python scripts. In order to protect user identity in both applications, non-attributable aliases were substituted for usernames. These aliases were cryptographic hashes of the usernames combined with a random string.

Many of these modifications and the implementation of the xAI tool were possible because both the object scanning and incident tracking applications are custom-developed tools. In many organizations, this may not be the case. However, most commercial applications used for both object scanning and incident tracking will maintain similar audit records about user interactions with the system. Therefore, it is not unreasonable to expect that most user interactions captured in this experiment will be reproducible in various commercial products.

3) Subjective trust and explainability usefulness: Survey instruments are a popular method for collecting subjective data. xAI methods claim to increase a user's trust in an AI model. To understand the trust level of and satisfaction with explanations from xAI, we used two existing scales to measure analyst perceptions: the Trust Scale recommended for xAI and the xAI Explanation Satisfaction Scale [6].

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 comprised of these two scales after at least one week of working with the xAI tool.

One known limitation of surveys is low response rate, which is exacerbated in operational settings in which human operators have a primary task to do. Filling out a survey requires ceasing the primary task, pivoting to the survey method (e.g. paper, digital website), and taking time to complete the survey. Our attempt to collect survey data was almost completely hindered by these environmental factors, which will be discussed in Section III.

B. AI tools in a live security setting

We identified important attributes to provide some context about the operational cybersecurity environment, specifically about the use of incident tracking and object scanning systems. From an enterprise perspective, there is high risk and associated cost of undetected malware. Security systems are tuned to be sensitive towards indicators of malware because the cost of undetected malware can be extremely high [20]. Security systems include multiple, sometimes partially overlapping, alerting criteria. Within this context there is a bias towards hard cases; easily detectable malware is automatically mitigated with existing tools and, therefore, not triaged or elevated for more investigation. Samples triaged by analysts are harder to classify and often involve contradictory predictions from competing (and highly accurate) mechanisms. 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 (IDS) [15]. In other words, the interpretation of feature space is not self-apparent (compared to some image classification problems) [17]. Notably, in the team we studied the individuals who triage alerts are largely disjointed from individuals who maintain the AI models (i.e. model maintainers).

It was 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 [5], 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 contradicted initial assumptions about how the tool would be used. These factors included:

a) Decision task and alternate decisions support paths: The analysts use the classification output from the AI model along 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 both before and after the tool was deployed to see if including explanations changed analyst behaviors. We sought to capture measurements such as average response time and whether explanation text sections are expanded to be read. We also made the visual presentation of explainability more appropriate compared to previous versions, which did not organize or present explanations in ways that could be quickly utilized during the decision making process.

b) Workflow: Much of the information that cybersecurity analysts use to make a triage decision exists in a central incident handling system, with little navigation required within the dashboard to find decision-critical information. This is a standard workflow in cybersecurity operations settings. For our study, this included the AI model classifier output (e.g. malicious, benign, uncertain), but not the xAI tool. User workflow should be considered prior to deployment of xAI techniques in some fashion to understand potential friction points for adoption.

c) Tool separation/location: The xAI tool exists outside the main dashboard where analyst conclusions are registered; it is located in a supporting program (i.e. the object scanning system) which required pivoting for engagement. While this

TABLE II. DATA COLLECTED FROM 17 ENTRIES

Period	Opened Explanation	Classifier Prediction	Analyst Agreed with Classifier	Time Event Open (min)	Notes	Total Event Views	Analysts with Entries Count
Pre-Tool	N	malicious	Y	0.50		19	1
Pre-Tool	N	benign	Y	30554.9	21 days	31	1
Pre-Tool	N	benign	Y	39.9		22	1
Pre-Tool	N	benign	Y	11.7		21	1
Pre-Tool	N	benign	Y	217.8		27	0
Pre-Tool	N	malicious	Y	35.5		21	1
Pre-Tool	Y	benign	Y	120.7		60	2
Pre-Tool	N	benign	Y	9.9		32	0
Pre-Tool	Y	malicious	Y	48.3		18	0
Post-Tool	Y	uncertain	N/A	52.0		43	1
Post-Tool	N	malicious	Y	0.87		11	0
Post-Tool	N	benign	Y	5.55		10	0
Post-Tool	N/A	benign	Y	8.3	object scanning system was not opened	77	1
Post-Tool	N	benign	N	N/A	Still under investigation/no close time	85	1
Post-Tool	Y	benign	Y	1395.7	1 day	27	2
Post-Tool	N	benign	Y	7.55		16	1
Post-Tool	N	malicious	Y	79.0		26	2

object scanning system is routinely accessed by analysts, the addition of the tool was not immediately obvious. To mitigate, 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.

d) Number of end users and their roles: In our scenario, there is a different 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 included all users who interact with the xAI tool, not just the incident responders who are primarily responsible for handling incidents in a given week.

III. RESULTS

We collected log data from the system with and without xAI tools without interrupting the incident handling task with and without xAI tools. We then measured user trust and perception of usefulness of the xAI tool. This section describes our findings and challenges with obtaining useful data in cybersecurity settings.

A. Log data findings

As described in Section II, quantitative 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. Data were collected for 36 days pre-tool implementation and 43 days post-tool implementation. A total of 2834 unique alerts triggered by the classifier were included in the data. Of the 2834 alerts, 17 were promoted to events; these are the focus of this analysis (Table II).

One hypothesis was that the availability of a novel explainability tool would increase likelihood of an analyst seeking out an explanation from AI models. Surprisingly, we discovered that users rarely interacted with the xAI tool, even after training. As shown in Table II, out of 9 promoted events that occurred pre-explainability tool, the existing explanation tool was opened 2 times (22%). Of 8 events post-explainability tool, the new xAI tool was opened 2 times (25%).

TABLE III. TIME TO CLOSE EVENTS WHERE EXPLANATIONS WERE VIEWED.

Period	Time Event was Open (min)	Average Time Event was Open
Pre-Tool Pre-Tool	120.7 48.3	85
Post-Tool Post-Tool	52.0 1395.7	724

Why were the analysts not opening the explanations very often, even *prior* to the implementation of a novel explainability tool? A shift in thinking allowed us to appreciate the key finding in the pre-deployment data: *the targeted analysts did not use xAI in their daily workflows*. Further investigation helped us understand that the information sources and cues that analysts primarily used for their decision were located in other available tools, mainly the incident tracking system. Moreover, we concluded that salience of a new tool further decreased the likelihood that users engage with the tool, and a training intervention to overcome that limitation is an insufficient strategy.

Another hypothesis was that the introduction of the explainability tool would change the length of time it took for an analyst to make a decision to promote an alert to an event or ignore/close an alert. When investigating this hypothesis, we considered only the events where the explainability tool was opened and found that the average length of time (in minutes) for the 2 events existing tool was 85 minutes, whereas for the 2 events post-implementation of the xAI tool, the average time was 724 minutes. We note that one event post-tool was open for more than one day (Table III) and time comparisons may not be reliable with such a small sample size. However, the system instrumentation could theoretically be run for much longer, providing data about time-based metrics about the incident life cycle as well as insights about how the tool is used and adopted over time.

TABLE IV. ANALYST AGREEMENT WITH CLASSIFIER IN EVENTS WHERE EXPLANATIONS WERE VIEWED.

Period	**Classifier Prediction**	**Analyst Agreed with Classifier?**
Pre-Tool	benign	Y
Pre-Tool	malicious	Y
Post-Tool	uncertain	N/A
Post-Tool	benign	Y

TABLE V. UNIQUE ANALYSTS WITH ENTRIES FOR EVENTS WHERE EXPLANATIONS WERE VIEWED.

Period	Total Event Views	Analysts with Entries Count	Average Unique Analysts with Entries
Pre-Tool	60	2	2
Pre-Tool	18	0	
Post-Tool	43	1	1.5 (2)
Post-Tool	27	2	

B. Subjective data findings

As described in Section II, qualitative data were collected in an online survey sent to $n = 11$ analysts. Questions probed user trust and perception of usefulness of the xAI tool [6]. Unfortunately, we had only one analyst complete the survey within the study time period. This was likely due to pressures from the operational environment and the intrusiveness of surveys during normal operations. Analysts simply did not have the bandwidth or time to complete surveys. Due to small sample size we were unable to analyze our instrumented event data by an average trust score. This result (or lack thereof) reinforces the need to reduce reliance on intrusive methods to circumvent issues with interrupting analysts' tasks.

C. Deployment challenges

We thought that an analyst's rate of compliance with the model output might change depending on whether they were using their old tool or the new tool to view the model explanation. In this case, of the four events (pre- and post- tool period), the analysts agreed with the classifier output, and in the fourth instance the classifier's output, was uncertain (Table IV). We found no difference in compliance between pre- and post-tool deployment.

With respect to an event, multiple analysts can add information into the event log via the incident tracking tool. We wondered if events where an explanation was viewed would have a different number of unique analyst entries. In Table V, we show that there is no difference. Finally, we thought that analysts might view an event more often if an explanation was opened. We found very little difference in the number of event views during the pre-tool phase versus the post-tool phase (Table VI).

IV. DISCUSSION

The results indicate that the xAI tool was not used by analysts in live cybersecurity operations; rather information and cues from other tools were used to support decision-making. The explanation capability was added to the existing system with the assumption that understanding model rationale would help with incident response 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 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.

TABLE VI. NUMBER OF TOTAL VIEWS FOR EVENTS WHERE EXPLANATIONS WERE VIEWED

Period	Total Event Views	Average Event Views
Pre-Tool	60	39
Pre-Tool	18	
Post-Tool	43	35
Post-Tool	27	

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 resolved. However, we believe that injecting a new xAI tool into existing models that analysts *already trusted* impacted our ability to detect any gain in confidence and reduction of skepticism. Essentially, we were unable to understand if analysts trusted the xAI tool because it was confounded with trust in the larger system already in use.

A. Challenges and lessons learned

Though the deployment of our tool was not considered successful in cybersecurity incident response, we share challenges and lessons learned with hopes of informing other research with considerations that can mitigate risk of failure. Cyber-related lessons are located in Table VII, while AI deployment lessons are located in Table VIII.

We faced several challenges in deploying the tool in a live setting. First, due to location of the xAI tool, which was embedded in an accordion menu in a supporting system, we expected a relatively low level of 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. Thus, we suggest carefully considering the task flow in terms of software tools being used before deciding on location of the new tool. Pilot testing can help identify potential issues with user interface designs and ease of access.

Our lessons learned also inform efforts for deploying AI models specifically "in the wild" that might be useful for others developing xAI for use in real-world settings. The study originally aimed to conduct more fully controlled field experiments, which quickly evolved into tool improvement. Over the period of about six 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.

TABLE VII. CHALLENGES IN DEPLOYING TOOLS IN CYBERSECURITY

Challenges	**Considerations**
Skilled cyber operators are highly skeptical of new tools due to the high consequences in security. Yet, once operators gain trust in a tool or resource, they develop robust workflows based on those tools and information sources.	Injecting new tools or features into existing workflows can be challenging if trying to detect usefulness and usability. Developers should carefully consider the location of the tool and validate the usefulness in a smaller, more controlled pilot. Concept testing could help mitigate wasted time and resources in developing a tool or feature that may not add value to the analyst's workflow.
Triage tasks in cyber operations' time-sensitive environments.	Introduction of new tools (which take time to learn) or data collection methods that impact time-based metrics of performance should be minimized or avoided whenever possible. System instrumentation that allows for non-intrusive data collection can provide some valuable insights about how tools are used while also capturing time-related data.
Cyber operations' technologically complex environments, with wide range of tools and data sources required to complete daily tasks.	New tools should aim to reduce complexity of the task and/or environment in cybersecurity. System instrumentation can be tricky across multiple tools, especially if a given system includes proprietary code. Different scripts and even redundancy may be needed to reliably capture data at the appropriate resolution.

TABLE VIII. PRACTICAL CONSIDERATIONS FOR XAI DEPLOYMENT

Practical Considerations	**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?

We validated previous findings with related studies [18] that the task context in which the xAI tool was deployed dictated how much it might actually be used and for what purposes. Though expected, we were still surprised by some aspects of context that ultimately 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. However, we reiterate that these contextual factors did affect our ability to collect certain types of data. Based on the above lessons learned, we offer considerations for developing and deploying xAI tools in live contexts. The questions in Table VIII can help guide decisions and mitigate risks during various stages of development and technology transfer.

Second, we anticipated challenges related to the environment in which the xAI tool was deployed. Though we found this to be less relevant for our use case, contextual factors (e.g. time pressure, task scope, etc.) may impact its usefulness and adoptability. Cybersecurity incident responders are known to experience high alert loads and are subject to different kinds of cognitive biases when interacting with intrusion detection systems [7]. These settings have a history of high turnover and burnout [2], [13], and judgments about alerts are often made with pressure from long a queue of alerts or time expected to make a decision [11]. 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.

Third, we learned that our user base seemingly trusted the output of the AI model to the extent that they did not explore novel explainability tools, similar to previous conclusions on non-cybersecurity users [18]. 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 activities related to validating the underlying AI model. However, this realization led us to identify two new research 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 explored the first question in a follow-on study, but future research should include exploring the second question, particularly ways of enabling contribution without interrupting normal workflow.

Finally, we recognize that the design and implementation of the tool could have contributed to its lack of use and adoption by the analysts. TreeSHAP is currently considered state-of-the-art in xAI, but has not been validated with user studies. In short, we assumed certain design principles and formats based on mathematical research in xAI, and that this xAI tool would help analysts focus manual analysis work.

B. *Follow-on study of secondary user group*

Findings from our study determined that AI model maintainers, or experts tasked with training AI models and monitoring their performance, are more invested in verifying model outputs to improve model accuracy than cybersecurity analysts. Within the organization included in this study, model maintainers had deep expertise in AI technologies in addition to cybersecurity. Accordingly, we pivoted our efforts to understand this user base better. Our exploratory research objective was to understand model maintenance tasks, information needed to evaluate model performance (specifically outputs of the xAI tool), and in doing so identify key differences from the cyber analyst workflow. The results of the follow-on study indicated that model maintainers need different information than incident responders for their respective tasks.

We engaged this small group of people through interviews to capture their task goals for model maintenance and their perspectives on the xAI tool. We found that the roles of these individuals vary greatly. To capture a wide variety of potential factors, we conducted one-hour semi-structured interviews with the model maintainers (n = 3) to understand roles and goals of each participant. The interview protocol is included in the appendix. The qualitative interview data was analyzed by one researcher with experience in qualitative coding and analysis. Statements from the participants were identified by thematic interest and summarized in the findings. Due to low sample size and high variation across the sample, we present this research as useful insights for current and future work as it pertains to AI models in practice cautioning that our research findings should not be interpreted as generalizable beyond the operational environment from which they were collected.

1) Roles of model maintainers: Each of the participants had different roles when interacting with an AI model of interest. The first two individuals had supporting roles that aid incident responders when needed, improve the model through identification of specific samples, and help monitor the model's performance. In addition to this support, both individuals had unique roles supporting at least one additional facet. The first individual had deep knowledge of underlying AI models and had a major role in creating them. This knowledge helps this person identify specific samples for model maintainers to consider, ensuring coverage of new and emerging threats. The second individual had some knowledge of the underlying models, provided support to the model maintainers, and also filled an analyst role. The purpose of this role is to help make the model more usable to analysts. This person's prior experience as an analyst helps in understanding how AI inference can be used to make decisions for specific observations. The primary model maintainer (third individual) developed classifier algorithms, identified new features, built processes for training and testing, deployed AI models, and performed model maintenance at regular intervals (weekly/monthly).

Model maintenance and retraining is continuous and indeterminate in length as the model maintainer continues to add new samples to the training set. This is especially true in cyber defense where attacks are constantly evolving. The model improvement process happens in 3 phases. First, the maintainer must find incorrect predictions and edge cases, conducted through a case by case review. Cases can be identified by supporting roles and analysts or through the maintainer's own queries and analysis. Next, the model maintainer must judge if particular samples are benign or malicious, sometimes redoing incident analysis and collaborating with others to reach a conclusion. Like the analysts, the model maintainer relies on contextual information about the observation/case. This contextual information is extraneous to the model itself. This could include the actual artifact (email, PDF, etc.) and summary data from supporting systems (e.g., Splunk) but varies by case. Last, the model maintainer must decide if the case is relevant to improving the model and if it should be added to the training set, doing error analysis, and updating model parameters.

2) Model goal and health metrics: The goal of the AI model, and thus of the model maintainers, is to detect threats reliably and consistently. Complete detection is not realistic given the dynamic, ever-changing nature of cyber threats and the static nature of most ML models. Despite this goal of deception by malware authors, there are some quantifiable metrics that are used to help gauge model performance – namely, model confidence should increase and incorrect classifications should trend to zero. Incremental improvements over multiple iterations of the model is desired. This is achieved by adding new observations that train the model to detect new and unique threats. The model itself is meant to assist a human in making a judgment about a given observation. Thus, two *effectiveness* metrics are linked to analyst interaction. The model should indicate when the model has low confidence in a prediction to draw human attention, and the tool should monitor how often analysts look at the explanations as a potential indicator of usefulness.

3) Key aspects of the model and its outputs: The following points are key aspects of the model such that it can meet the decision making needs of both analysts and model maintainers. Note that these model characteristics and outputs apply both to individual observations (as seen in the analyst workflow) and at the overall model level (as seen in the model maintainer role).

Confidence / Certainty of prediction: This is perhaps the most important piece of information beyond the prediction itself for the analyst and model maintainer to judge the output of the model. All participants noted this was missing from the current explainability tool and that it would be difficult to critically evaluate the model output without it.

Classification accuracy: Was the model correct? Accuracy requires knowledge of ground truth (currently determined by a human retroactively) but could be a later addition by a human user to help evaluate model performance.

Feature filters: The version of the xAI tool we tested included the top 10 features that contributed to the model prediction. However, the top 10 features only show a small portion of the values that contribute to the overall prediction. It may be helpful to have the option to see all features (or at least know the total number of features used in a given instance) or a user-constrained set of features such that they have appropriate framing for the judgment.

Total values for benign/malicious/overall: The top 10 shortlist of features can be somewhat confusing. Analysts are expected to evaluate the prediction against numeric outputs of the model, but due to the number of features it is unrealistic to show all features and feature values. Alternatively, the model could show the total values for benign and malicious-predicting features and the net value to share with the analyst how ”close” the prediction was to the center. The total magnitude in either direction would help analysts understand how close the prediction was to ”uncertain”.

Feature definitions: Currently, feature names are somewhat obscure if the user does not have knowledge of the model's architecture and function. For instance, the feature name “email_to_domain_other” is somewhat easy to parse as “domain” is ambiguous referring to either internal or external, whereas “pdf_text_keyword_view” is more challenging to decipher with multiple potential meanings. Model maintainers have this knowledge, but analysts may not. It is important to give this meaning to a user such that they can properly interpret it and make judgments. Features that are useful for machine learning may not have semantic meaning to a human.

Raw feature values (pre-normalization): These data would indicate feature-specific numbers, which may not be useful in all contexts. Not all users said this would be important, especially if the user is not familiar with the distribution of each feature.

Sparkline of the distribution: This small visual would show where the observation falls in the distribution for that feature. This addition could complement the raw values as contextual information to help decision making.

Global feature importance: This would be the same set of numerical values for the set of features included in the model. Global feature importance measures the importance of the feature for the entire model. It indicates how much impact that one feature, out of hundreds, has on a classification outcome. Building on the previous example, perhaps the feature checking whether email domain is internal versus external would be a significant contributor to a classification outcome in this model checking for malicious emails. The global feature importance would not change observation-to-observation unless or until the model is retrained.

Class feature importance: This is different for each observation. Local feature importance measures the contribution of the feature for that specific observation. For instance, observation k was run through the model, where the observation is an email that has an attached image. In this instance it is possible for "pdfstructure_image_dimensions_len" to be a bigger contributor to the prediction (model outcome) than average or than other observations. The set of numerical values for the set of features used in this observation would vary per observation.

Information that triggered the feature value (context): For instance, if a particular feature strongly contributes to the prediction of "malicious" the feature itself might not give enough information for the analyst to judge the outcome. The analyst may want to know what exactly caused the feature to produce its result. This information is not currently included in or accessible through the explainability tool. Currently, analysts must pivot to the actual observation and its artifacts (e.g. email, attachments, etc.) to find this information. Reducing the need to pivot between tools would save the analyst time while also increasing confidence in the tool.

4) Cross-cutting use cases for xAI: Based on statements from participants, the following examples demonstrate how the xAI tool might be used in both the analyst and model maintainer contexts.

1) **Evaluating residuals:** Being able to see and explore the residuals would help in determining if new features are needed.
2) **Hypothesizing new features:** Being able to see the model from a higher level; knowing what features are already included and what their respective coverage is would help in hypothesizing if/which features should be added to the model.
3) **Finding similar examples:** It is helpful to be able to evaluate an observation against similar records. Helping an analyst or model maintainer identify those observations and pivot from the tool would help improve this comparison process.
4) **Specialized/custom queries:** Analysts and especially model maintainers expressed interest in being able to view multiple observations and control those outputs using specialized or custom queries.
5) **Investigate specific aspects of an observation:** It would be helpful to identify an aspect of the explanation to draw the analysts' attention to something specific. For instance, if a feature indicates that there is something present (or missing) in the email or PDF that indicates a potentially malicious artifact, providing that information (i.e., what is the actual evidence observed by the model) to the analyst would improve their confidence and efficiency in evaluating the classifier output.

5) Other notable points from the study campaign: Another finding was that participants from all three studies indicated a trust dynamic that might not match how we are thinking about the problem. The main user who performs model maintenance indicated more trust in the analyst's decision than the model, but our studies revealed that the analysts are likely to agree with AI-driven predictions. This creates a strange paradox of trust in these settings, both between humans and between humans and AI.

We found that the explainability interface has low generalizability to other models and processes within cybersecurity operations, but the same key aspects we identified may apply in xAI design. Other AI models are employed, but they are commercial models embedded in purchased tools and are not within the control of model maintainers to train and prune. Practically speaking, there are many (raw and meta) data associated with events that analysts evaluate, and we noted challenges in the evaluation and training processes that could be addressed without AI or explainability. For instance, when viewing an event flagged by a classifier, analysts are interested in identifying what portions of a PDF caused the classification. However, this is not immediately obvious, and the existing tool/platform does not tie back to the original PDF but rather simply provides a list of features. Moreover, features in the model that contribute to "maliciousness" do not provide evidence in the same window, and the analyst is forced to pivot in order to properly evaluate the feature and corresponding information that triggered it. These small improvements in usability would help by improving the confidence and efficiency of incident response analysts.

V. Conclusions

While conducting a study aimed to understand benefits of xAI tools in cybersecurity operations, we learned that analysts seemingly trust the output of the AI model within the context of their current tool set and do not explore provided explanations. Rather, existing tools are used to validate the output of AI models. In this context, the output from the AI classification model was embedded in cybersecurity analysts' main workflow while the new xAI tool was not.

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

Acknowledgment

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. This work was supported by the Laboratory Directed Research and Development pro-gram at Sandia National Laboratories. 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. Content from SAND2021-11358.

References

[1] A. B. Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. García, S. Gil-López, D. Molina, R. Benjamins *et al.*, "Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai," *Information Fusion*, vol. 58, pp. 82–115, 2020.

[2] J. P. Bourget, "Addressing analyst fatigue in the soc," https://www.brighttalk.com/webcast/288/224207, 2016.

[3] B. J. Erickson, P. Korfiatis, Z. Akkus, and T. L. Kline, "Machine learning for medical imaging," *Radiographics*, vol. 37, no. 2, pp. 505–515, 2017.

[4] R. S. Gutzwiller, S. Fugate, B. D. Sawyer, and P. Hancock, "The human factors of cyber network defense," in *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, vol. 59, no. 1. SAGE publications Sage CA: Los Angeles, CA, 2015, pp. 322–326.

[5] R. R. Hoffman, "The concept of a "campaign of experimentation" for cyber operations," *The Cyber Defense Review*, vol. 4, no. 1, pp. 75–84, 2019.

[6] R. R. Hoffman, G. Klein, and S. T. Mueller, "Explaining explanation for "explainable ai"," in *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, vol. 62, no. 1. SAGE Publications Sage CA: Los Angeles, CA, 2018, pp. 197–201.

[7] A. Lemay and S. Leblanc, "Cognitive biases in cyber decision-making," in *Proceedings of the 13th International Conference on Cyber Warfare and Security*, 2018, p. 395.

[8] S. M. Lundberg, G. Erion, H. Chen, A. DeGrave, J. M. Prutkin, B. Nair, R. Katz, J. Himmelfarb, N. Bansal, and S.-I. Lee, "From local explanations to global understanding with explainable ai for trees," *Nature machine intelligence*, vol. 2, no. 1, pp. 56–67, 2020.

[9] L. A. McNamara, "Adoption challenges in artificial intelligence and machine learning:why technology acceptance is hard (and what we can do about that)." 1 2020. [Online]. Available: https://www.osti.gov/biblio/1763870

[10] M. Nyre-Yu, "Determining system requirements for human-machine integration in cyber security incident response," Ph.D. dissertation, Purdue University Graduate School, Oct 2019.

[11] C. Petersen and R. Lentz, "Surfacing critical cyber threats through security intelligence: A reference model for it security practitioners," *SANS Institute*, 2015.

[12] E. Raff, J. Barker, J. Sylvester, R. Brandon, B. Catanzaro, and C. K. Nicholas, "Malware detection by eating a whole exe," in *Workshops at the Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[13] C. Richmond and P. Lindstrom, "Idc security survey: As the job churns," IDC Corporate USA, Tech. Rep., 2015.

[14] R. Ruefle, A. J. Dorofee, D. A. Mundie, A. D. Householder, M. Murray, and S. J. Perl, "Computer security incident response team development and evolution," *IEEE Security & Privacy*, vol. 12, pp. 16–26, 2014.

[15] M. R. Smith, N. T. Johnson, J. B. Ingram, A. J. Carbajal, B. I. Haus, E. Domschot, R. Ramyaa, C. C. Lamb, S. J. Verzi, and W. P. Kegelmeyer, "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*, 2020, pp. 49–60.

[16] C. Smutz and A. Stavrou, "Malicious pdf detection using metadata and structural features," in *Proceedings of the 28th annual computer security applications conference*, 2012, pp. 239–248.

[17] R. Sommer and V. Paxson, "Outside the closed world: On using machine learning for network intrusion detection," in *2010 IEEE symposium on security and privacy*. IEEE, 2010, pp. 305–316.

[18] M. C. Stites, M. Nyre-Yu, B. Moss, C. Smutz, and M. R. Smith, "Sage advice? the impacts of explanations for machine learning models on human decision-making in spam detection," in *International Conference on Human-Computer Interaction*. Springer, 2021, pp. 269–284.

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

[20] M. Willett, "Lessons of the solarwinds hack," *Survival*, vol. 63, no. 2, pp. 7–26, 2021.

Appendix: Follow-On Interview Questions

1) Please describe your job role(s) as it pertains to model maintenance.
2) What is your primary goal when performing model maintenance? How do you know when you have achieved it?
3) How do you normally interact with the machine learning (ML) models / outputs? About how often?
4) What other machine learning models do you interact with in this way?
5) What information would you normally be searching for when you're investigating the ML output (e.g. classifier = malicious/benign)? What questions do you normally ask in your head as you're doing this? ——*(Show xAI representation to participant)*——-
6) When you see this presentation, what do you think it means?
7) What do you think the features represent? The feature values?
8) What do you think the direction of the bar represents?
9) What do you think the color represents?
10) Given this visual, can you imagine any difficulties in obtaining the information you need?
11) Is the amount of information presented appropriate for your needs in evaluating the model output as a model maintainer? If not, please describe.