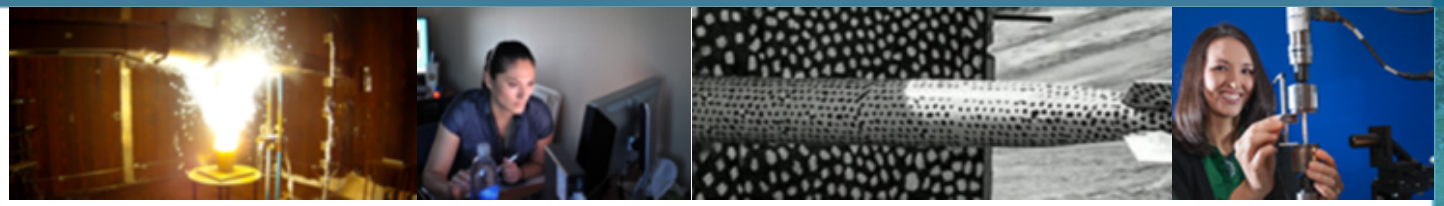
SAND2022-5301C

Sandia National Laboratories

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



Elizabeth S. Morris, Sandia National Laboratories
Albuquerque, NM, USA

April 28, 2022
Emerging Themes Session
Usable Security and Privacy (USEC) Symposium 2022
San Diego, CA, USA

U.S. DEPARTMENT OF ENERGY | NNSA

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.

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.

Paper & Authorship

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**

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.

Bottom Line: Lessons for Cyber Usability Research

1. Designing and deploying new tools in cybersecurity contexts …
   - An explanation capability for an existing AI tool did not help with incident response triage tasks in a live cybersecurity operation
   - Developers should carefully consider location of the tool in a smaller pilot study
   - Concept testing could mitigate wasted time and resources developing a tool that is not value-added for a cybersecurity analyst's workflow
   - New tools should reduce complexity of the task and/or environment in cybersecurity
2. Instrumented data collection methodology …
   - System instrumentation that allows non-intrusive data collection can provide valuable insights about how tools are used while also capturing time stamps
   - Instrumentation is difficult across multiple tools; different scripts and even redundancy are needed to capture data at the appropriate resolution in these environments

Explainable AI – A Brief Overview

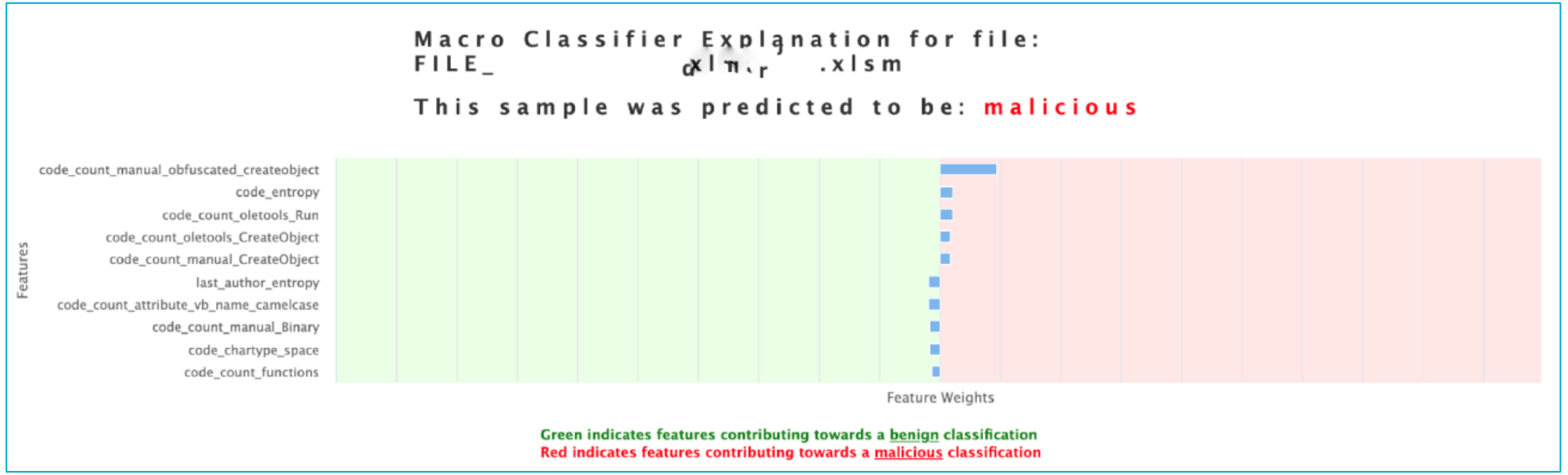
- Explainable AI (xAI) refers to how well a machine learning (ML) model's output, and especially the "rationale" behind its output, is understood by the human user
  - A classification tool could simply present to a cybersecurity analyst "malicious" or "not malicious"
  - OR, a classification tool could present the "malicious" or "not malicious" output with supporting visual information so that the cybersecurity analyst better comprehends why the conclusion (output) was reached by the underlying model
- Impetus for xAI: "Black box" models (neural networks, support vector machines) that are not easily understood by humans
  - Engineering & AI developer objectives: Ensure efficacy, improve control, and progress model performance
  - End user objectives: Understand context of explanation, communicate uncertainty, enable user interaction with explanation

Artificial Intelligence in Cybersecurity

- Artificial Intelligence (AI) algorithms may improve cyber infrastructure defense
  - Malware detection
  - Malicious pdf detection/code vulnerabilities
  - Phishing detection

- Cybersecurity analysts, a.k.a. “incident handlers”, may be skeptical of automated tools
  - Analysts have access to a variety of tools to investigate flagged events
  - False negatives can be extremely high-impact

Explainable AI for Cybersecurity Tool



- Our goal: Evaluate the effectiveness and efficiency of ML model output in an incident handling environment before and after an xAI tool was introduced
  - Instrumented system (log) data collected pre- and post-xAI tool deployment
  - Survey data on trust & confidence in the xAI tool, and satisfaction with the xAI tool
- Our main finding: Incident handlers rarely interacted with the xAI tool both pre- and post-deployment

The xAI for Cybersecurity Tool … Outcomes

- Deploying the xAI tool was considered a failure due to lack of use by incident handlers
- Subjective trust and explainability usefulness was unable to be measured due to low response rate from incident handlers (low response rate exacerbated in operational settings)
- Pivoting for engagement with the xAI tool (due to the xAI tool location) may have reduced analyst use of the tool
- Existing tools are used to validate the output of AI models
- AI model maintainers are more invested in verifying model outputs than cybersecurity analysts (or incident handlers)

Practical Considerations for xAI Deployment

- Who are your end users?
  - Who uses the model outputs, and in what way?
  - How does the xAI tool help users 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?

Practical Considerations for xAI Deployment *(continued)*

- 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?

Lessons for Cybersecurity Usability Research

- Concept testing could mitigate wasted time and resources developing a tool that is not value-added for a cybersecurity analyst's workflow
- New tools should reduce complexity of the task and/or environment in cybersecurity
- System instrumentation that allows non-intrusive data collection can provide valuable insights about how tools are used while also capturing time stamps

Questions?