Assignment: Research Proposal

Name: Cristhian M. Faria-Sanchez

Student ID: 12693874

Class: Research Methods and Professional Practice April 2025 B

Research Proposal Presentation: The Augmented Analyst (Transcript for presentation)

Welcome to my presentation, my name is Cristhian M. Faria-Sanchez, and today I'll be presenting my research proposal for my potential Master's capstone project. The title of my proposed research is: “The Augmented Analyst: A Simulated Study Evaluating the Impact of Explainable AI on Security Operations Centre Workflow”. This project relates to the CyBOK 1.1 Knowledge Areas of “Human Factors(4.x)” and “Security Operations & Incident Management(8.x)”.

Significance & Research Problem

As many professionals in the field know; Security Operations Centres were and still are at a volatile point. They're grappling with what the literature calls a “chronic crisis”, a deluge of data that leads directly to analyst burnout. We're asking our defenders to find a needle in a haystack, but that haystack is growing exponentially every day, leading to significant financial and security repercussions for organizations.

Artificial Intelligence(AI) powered security systems arose as the definitive solution. Through varied uses of the technology, individual analysts and companies alike have applied AI techniques to bolster defenses, allowing for the analysis of large data sets, discovery of anomalies and moving past the limits of traditional signature reliant security systems.

In the field of endpoint security alone, Machine Learning(ML) assisted endpoint security systems have made malware based intrusions almost a thing of the past. As highlighted by Crowdstrike’s 2025 Global Threat Report, advanced persistent threats (APT’s) have shifted their focus from malware/vulnerability based instructions, to identity based attacks(Crowdstrike, 2025). The report goes further and states that 79% of attacks reported were without use of malware. Experts in the field believe that this is in large part, due to EDR/ML advancements making exploit based intrusions too costly for attackers, so they have shifted their focus to “lower hanging fruit”.

The advent of this “Miracle Solution”, however, is not without its own unique challenges. As recent research from Greshake and colleagues has shown, AI models can be fundamentally untrustworthy as they are vulnerable to novel attacks like ”indirect prompt injection”, and too commonly are incorrect due to “hallucinations” or fundamental “misunderstandings” based on questions being asked that are not the most common type in its training .

An analyst simply cannot afford to act on a recommendation without understanding the “why” behind it. This isn't just a matter of trust; it's a matter of operational security and a matter of responsibility in the realm of risk. On the other hand, if a Machine Learning(ML) enabled machine makes a detection due to what the system considers “Anomalous Behavior”, an analyst is often required to act on these alerts and having to rely on the system to explain why this behaviour is “Anomalous” or be left trying to find this information themselves, defeating the purpose of the AI’s inclusion. This leads into a fundamental paradox where AI can simultaneously be the solution, as it is able to detect the “needle in the haystack”, but also being a source of extra burnout and risk responsibility to an analyst if it is not able to fully explain the reasoning and relevancy behind its detections.

There appears to be a need for empirical studies that measure the real-world impact of Explainable AI (XAI) on the daily workflow of a security analyst. My project will aim to contribute to the discipline by providing quantitative and qualitative data on the current issues presented by an “AI augmented” investigation process and whether the implementation XAI functionality can assist in mitigating the downsides of the AI augmented investigation process.

Research Question, Aims & Objectives

This leads me to my central research question: Can the integration of Explainable AI (XAI) into an analysts workflow improve the performance, efficiency and confidence of a SOC analyst during a live-fire attack simulation? To answer this question, I've established three primary objectives, each designed to evaluate a different dimension of issues an analyst may face.

1. **To Quantify Performance:** My first objective is to measure the impact of XAI on key performance indicators. Specifically, I will track the Mean Time to Detect and Mean Time to Respond to a simulated threat, which are industry-standard metrics for SOC efficiency.

2. **To Evaluate Decision-Making:** Second, I will assess how XAI influences the quality of an analyst's decisions. Using a non-technical skills model, I will evaluate the accuracy of their findings and the soundness of their response choices.

3. **To Analyse the Human Factor:** Finally, and perhaps most importantly, I will gather data to understand how XAI affects the analyst themselves. I will measure their cognitive load using the NASA Task Load Index and qualitatively assess their situational awareness and trust in the automated system.

Key Literature & Theoretical Framework

Binbeshr et al. (2025) propose that the future of organizational security is through the development of Cognitive Security Operation Centers (C-SOC's), where analysts work alongside AI systems, in an "augmented intelligence" relationship. However, I will extend this by focusing on human-centric evaluation. The work of Zhang et al. on the CYBENCH framework highlights the need for reliable methods to evaluate AI capabilities. My project was inspired by this benchmarking philosophy, but crucially, I intend to benchmark the entire human-AI relationship, not just the AI in isolation.

To evaluate the human element, I will draw on standard testing methodologies. The NASA Task Load Index, developed by Hart and Staveland in 1988, will provide a robust, quantitative measure of cognitive workload. For qualitative insights into the decision-making process, I will adapt principles from the NTS-CS framework for evaluating non-technical skills in cybersecurity, developed with sponsorship. Ultimately, this research aims to move from a theoretical discussion of human-AI teaming to a practical demonstration of what Augmented Intelligence may look like in a live-fire attack.

Methodology Part 1 - The Technical Environment

I will employ a comparative experimental design within a controlled, virtualized lab environment. The majority will be hosted on a Proxmox Virtual Environment (essentially everything except the LLM, unless I can obtain the resources to implement a locally managed LLM). I chose this environment as it provides reliable snapshotting capabilities, allowing me to

reset the entire environment between experimental phases; which is in an attempt to minimize internal variables that may affect the project's results.

The lab itself will roughly consist of three core components:

- A Windows Desktop virtual machine, which will serve as the target endpoint for the simulated attack.
- A GNU/Linux Server VM(likely Ubuntu), which will host both the SIEM for security monitoring and the MITRE Caldera server for adversary emulation.
- An internet gateway VM that will act as a firewall, logging proxy and potentially; as a parser for logs towards the LLM API if this cannot be performed at the log forwarder on the linux system.

(note for professor: In the case where I can make a purchase a machine with an APU, such as the Framework desktop as the cost is not too high, I can also foresee this setup being performed where instead of an ubuntu machine, I can send all logs to the machine itself via a physical connection and have it handle the log intake, ML analysis and detection module, SIEM and the local LLM with the XAI for the context enrichment).

Log sources are critical for this experiment. I will focus on only the most important data sources (due to potential costs/system constraints). This will essentially involve parsing the following logs for those that would typically be recorded in an SME's SIEM: Sysmon logs, endpoint agent alerts, networking logs such as netflow/BPF PCAP, potential forensic artifacts, and potentially Event Tracing for Windows (ETW) logs. The intention is to provide a rich and realistic dataset for both the analyst and the AI to work with, while simultaneously being achievable within my capabilities and budget.

Methodology Part 2 - The Experiment

White the methodology is still being ironed out, I conceptualize that the experiment will proceed in two distinct phases for each participant, allowing for a somewhat direct comparison.

Phase 1: Baseline AI Assistance:

The analyst will investigate alerts enriched by a script that provides minimal, non-explanatory information, as seen with most current systems. For example, the AI might provide a severity score and a MITRE ATT&CK Tactic ID, alongside a generic description, but

no additional reasoning.

Phase 2: XAI-Enhanced Assistance:

The analyst investigates a very similar attack, but this time, the alerts are enriched by the XAI script that provides a narrative explanation. In theory, to facilitate this, I could use a cloud-based API, such as OpenAI's GPT-4o or Anthropic's Claude 3. Although, to make the project more in line with my goal, I will also explore the feasibility of using a locally-hosted model that may allow for the same system to create and explain the detection. This is the ideal situation, although less likely under my constraints.

Methodology Part 3 - Evaluation Frameworks & Challenges

- **Quantitative Performance:**

  I'll measure the Mean Time to Detect and Respond. Likely in the form of set goals, e.g. identifying the C2 channel; which will be done while I record.

- **For Cognitive Load:**

  I will administer the NASA-TLX, a standardized survey that quantifies the mental workload experienced by the analysts in each phase. I am currently considering allowing for short breaks between sections which will be where I will perform the survey.

- **For Qualitative Analysis:**

  I will conduct structured interviews guided by the UK Railroad Safety non-technical skills framework, focusing on the analysts' situational awareness and decision-making processes.

Potential Challenges

The following are challenges I anticipate will be relevant throughout this project:

- **Technical Complexity:** The primary risk is unforeseen technical issues within the lab environment or in developing this project due to potential lack of capability in developing this project on my end. In an attempt to mitigate the potential technical issues during testing, I scheduled for a pilot study in week 6 to identify and resolve any bugs before the main data collection begins. This will either be performed on my own or with an assistant. For the concerns of the technical requirement to create the testing, I will attempt to follow similar security “home lab” setups or recommendations by MITRE, and I will dedicate 3 weeks for the creation of the lab environment and testing plan.

- **Participant Recruitment:** While I am in contact with multiple security analysts who have expressed interest in participating in this project, I believe it is likely that I will experience issues in scheduling and finding sufficient participants. I will also have to decide on a requirement for entry such as accreditations/years of security experience, which is likely to narrow available participants.

- **LLM Unpredictability:** LLMs can 'hallucinate' or provide inconsistent output. This may affect both the detection and the explanations. I will attempt to manage this risk in three ways:
    1. Testing the exact detections in my pilot period.
    2. Making changes to make the systems more predictable such as having a 'temperature' setting of zero in the API calls to make the model's responses as deterministic as possible.
    3. Working to identify well trained models and XAI frameworks tuned for cybersecurity.

- **Cost/Time restraints:** I foresee potential challenges in both the potential time investment for this project, as well as the costs, both expected and not. Ideally, I can create the lab environment with some of my existing home lab equipment, however, lack of processing power can be a negative factor that will affect the project's legitimacy. The route of using cloud resources is feasible but has potential for large cost increases with scale. For this reason, I will attempt to speak with potential sponsors to gain funding for a testing equipment and cloud resources budget.

Description of Artefacts

With regards to the BCS requirements for a postgraduate project, I foresee multiple potential artefacts I can use. My artefacts may comprise various extracts from my simulated study environment and the structured methodology used. This includes the Proxmox lab configurations, SIEM configurations, Attack Simulation Scripts, the XAI enrichment script(or any changes made if I use an open source solution locally), and all resulting data from testing as discussed in the "methodology part 3" section.

Ethical Considerations & Risk Assessment

In consideration of necessary adherence to ethics and to minimize associated risk, I identified three main principles that I aim to follow throughout my testing:

1. **Informed Consent:** All participants will be fully briefed on the study's purpose, procedures, and their rights. They will be required to sign a consent form before participation.

2. **Anonymity and Confidentiality:** All collected data, from performance metrics to interview recordings, will be fully anonymized to protect the identity of the participants and their employers.

3. **Data Security:** All research data will be stored on an encrypted drive, accessible only by myself, to ensure its security and integrity.

The primary variable risk I identified associated with this study is the potential for performance anxiety among the participants. I plan to mitigate this by communicating that this is not a test of their personal skill, but rather an evaluation of the tools and processes. I will also let participants know that their anonymity will be a priority throughout the project.

Timeline of Proposed Activities

I have developed the following 16-week timeline, separated into six phases:

- **Weeks 1 to 4** will be dedicated to the setup of the lab and tools.
- **Weeks 4 to 7** will focus on the creation of the XAI script and creating the attack scenarios in Caldera.
- **Week 8** will involve a pilot study where I run through the scenarios on my own to identify potential issues
- **Weeks 9 to 11** are the main Data Collection window, where I will work with the three+ participants.
- **Weeks 12 to 14** will be for Data Analysis, where I will process the quantitative and qualitative data.
- **Weeks 15 and 16** will be for the Final Report write-up and project submission.

Conclusion

In conclusion, for this project, I aim to create valuable test data on the real-world utility of Explainable AI in a Security Operations Centres. By moving beyond theoretical discussion and into practical, human-centric evaluation, I aim to produce insights that are relevant and actionable for both academia and industry.

Thank you for your time.

References:

Binbeshr, F., Imam, M., Ghaleb, M., Hamdan, M., Rahim, M.A. and Hammoudeh, M. (2025) 'The Rise of Cognitive SOCs: A Systematic Literature Review on AI Approaches', IEEE Open Journal of the Computer Society, 6, pp. 360-375.

CrowdStrike (2024) CrowdStrike 2024 Global Threat Report. Austin, TX: CrowdStrike. Available at: https://www.crowdstrike.com/global-threat-report/ (Accessed: 10th July 2025).

Greshake, K., Abdelnabi, S., Mishra, S., Endres, C., Holz, T. and Fritz, M. (2023) 'Not What You've Signed Up For: Compromising Real-World LLM-Integrated Applications with Indirect Prompt Injection', in Proceedings of the 16th ACM Workshop on Artificial Intelligence and Security (AISec '23). Copenhagen, Denmark: ACM, pp. 79-91.

Hart, S.G. (2006) 'NASA-Task Load Index (NASA-TLX); 20 Years Later', in Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 50(9), pp. 904-908.

Hart, S.G. and Staveland, L.E. (1988) 'Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research', in Hancock, P.A. and Meshkati, N. (eds.) Human Mental Workload. Amsterdam: North-Holland, pp. 139-183.

Healey, C., et al. (2019) An NCSC-sponsored project to develop a taxonomy of non-technical skills for cyber security (NTS-CS). Available at: https://www.ncsc.gov.uk/files/NTS-CS-taxonomy-report.pdf (Accessed: 7 July 2025).

Malatji, M. (2025) 'Augmented Intelligence Framework for Human–Artificial Intelligence Teaming in Cybersecurity', Human-Centric Intelligent Systems..

Mohale, M. and Obagbuwa, I. (2025) 'A systematic review on the integration of explainable artificial intelligence in intrusion detection systems to enhancing transparency and interpretability in cybersecurity', Frontiers in Artificial Intelligence, 8. doi: 10.3389/frai.2025.1526221.

Mohsin, A., Janicke, H., Ibrahim, A., Sarker, I.H. and Camtepe, S. (2025) A Unified Framework for Human AI Collaboration in Security Operations Centers with Trusted Autonomy. arXiv:2505.23397. Available at: https://arxiv.org/abs/2505.23397 (Accessed: 7 July 2025).

Tian, S., et al. (2025) 'Exploring the Role of Large Language Models in Cybersecurity: A Systematic Survey', IEEE Transactions on Network Science and Engineering, XX(XX)..

Zhang, A. K., et al. (2025) 'CYBENCH: A FRAMEWORK FOR EVALUATING CYBERSECURITY CAPABILITIES AND RISKS OF LANGUAGE MODELS', in International Conference on Learning Representations (ICLR 2025