# Enhancing AI Transparency in Cybersecurity: Tackling the Black Box Problem through Explainable AI

CS4991 Capstone Report, 2025

Caroline Coughlin
Computer Science
The University of Virginia
School of Engineering and Applied Science
Charlottesville, Virginia USA
ghe9da@virginia.edu

## ABSTRACT

The black box problem in artificial intelligence poses an ethical barrier to innovation, as AI systems often produce outputs with limited or unclear explanations of their decision-making processes. During my internship at a cybersecurity company, I addressed the black box problem firsthand utilizing prompt engineering techniques, which optimized AI model responses and enhanced understandability. My work was centered on an AI application that scanned client documents and extracted answers for due diligence questionnaires, addressing crucial cybersecurity needs. By identifying discrepancies in AI outputs, refining input prompts, and implementing step-by-step explanations, I improved the model's accuracy and transparency, enhancing users' understanding of its thought processes. These contributions minimized the risk of inaccurate information, security vulnerabilities, and compliance issues while reinforcing trust in the AI model. Future work will examine additional Explainable AI (XAI) methods, such as Local Interpretable Model-Agnostic Explanations (LIME) and Shapley (ES) Values, to further narrow the gap between algorithmic transparency and practical application, ensuring AI models are both comprehensible and reliable in critical domains.

## 1. INTRODUCTION

In recent years, swift artificial intelligence advancements have led to the use of countless machine learning-based tools in everyday decision-making. These tools are designed to analyze vast datasets, uncover hidden patterns, and suggest solutions. While these tools make large tasks easier and add convenience to our lives, it is essential to simultaneously acknowledge the ethical consequences of implementing these systems in real-world scenarios. One major challenge with AI-driven solutions is the difficulty in understanding the rationale behind the algorithm's proposed solutions. Due to the existing opaqueness in AI algorithms, humans struggle to visualize and comprehend how deep learning systems arrive at their choices and projections [1]. Researchers in AI examine this issue within an explainability framework and have come to define it as the black box problem. An AI system is considered explainable if its operations can be simplified into an external representation that is comprehensible to humans [2].

In the case that the black box problem is not rectified, significant repercussions will inevitably follow, including a lack of accountability for AI-created decisions and errors or biases within data and models, leading to a decline in public trust. These issues additionally complicate the broader adoption of AI across various sectors. In legal scenarios, insufficient explainability may result in unjust outcomes that could have serious consequences, such as discriminatory sentencing. With regulations like the GDPR, AI systems that fail to offer explanations for their automated decisions may encounter legal challenges or sanctions [2]. Consequently, businesses in sensitive domains such as cybersecurity, healthcare, and finance may also experience penalties for utilization of ML models.

## 2. RELATED WORKS

Artificial intelligence possesses the capability of vastly improving cyber security and defense measures, thereby increasing stability in the cyber domain. As we analyze the contributions of AI from a systems lens, three areas of influence appear: system robustness, system resilience, and system responses [3]. In regards to system robustness, AI can assist in the verification and validation of software, eliminating tedious tasks, and increasing the efficiency of system testing. In addition, AI may improve system resilience with its pattern recognition capabilities as it applies to threat and anomaly detection. Furthermore, AI offers autonomous and semi-autonomous cybersecurity systems with integrated pre-determined responses to cyber attacks. Autonomous systems can additionally learn adversarial behavior and produce decoys to actively lure threat actors [3].

In order to harness the potential of AI in the cyber domain, we must mitigate the corresponding risks of usage. Areas of concern include data bias and fairness, privacy, accountability, transparency, job displacement and legal and regulatory challenges [4]. Algorithmic bias from skewed training data or designer prejudices may perpetuate systemic inequalities in AI models, detrimentally impacting certain demographic groups. When detecting anomalies in user behavior, AI may violate individuals' privacy and monitor a larger range of behavior than necessary. With advancements in machine learning-based AI models, various challenges arise in relation to computational complexity, data sparsity and model generalization [4]. The conjunction of rule-based AI and data-driven ML techniques may potentially generate a model of high accuracy and interpretability.

## 3. PROJECT DESIGN

The following section outlines the design process used to develop a transparent and reliable AI system for due diligence document analysis. The process is broken down into key requirements, components, and challenges faced during development.

### 3.1 Requirements

In order to adequately address the black box problem in the AI model utilized for due diligence questionnaires (DDQs), an organized approach centered on transparency, accuracy, and usability was critical. The main objective was to improve the explainability of the AI model's decision-making process while ensuring the extraction of precise responses from client documents remained high-performing. This called for enhancements in prompt engineering to better direct the AI's reasoning and response generation.

The system additionally needed to comply with cybersecurity best practices, safeguarding sensitive financial information from hedge funds and private equity firms. The project also necessitated an iterative testing approach to continuously refine prompt structures and assess model performance over time. Furthermore, the optimization of prompts maintained compatibility with the model's underlying natural language processing (NLP) framework. The model was constructed using Python and machine learning techniques for reference.

### 3.2 Key Components

The project entailed various components that worked in stages to advance the AI application's transparency and reliability. The first component was data extraction and citation analysis, in which the AI processed client documents and obtained relevant information providing answers to due diligence questionnaires. In this phase, extracted answers then had to be verified for accuracy and inconsistencies were to be detected in the AI's responses. Model responses were deemed inconsistent if logic was unsound or diverging answers were found in the corresponding client documents cited.

The second component in the project was prompt engineering, where specific, detailed instructions were designed to remedy inaccuracies and guide the AI in determining the correct answers in client due diligence documents. These prompts were tailored to be modular and scalable, enabling continuous, iterative refinement as the AI's behavior is monitored over time. To further combat the black box issue, an explainability layer was developed. This layer integrated an additional set of prompts into the model to generate step-by-step explanations behind each response provided. I designed specific instructions, detailing every logical step and reasoning that the AI should vocalize to the end user. As soon as I added these prompts to the model's knowledge base, it began to cite specific document sections and provide a sequential visualization of its thought processes behind each conclusion.

The final phase of the project entailed a testing and refinement cycle, where differing prompt versions were analyzed to determine which had the largest effect on AI transparency. The model's outputs were evaluated systematically and alterations to prompts were made based on observed weaknesses, ensuring continuous improvement in response accuracy and interpretability.

### 3.3 Challenges

Despite the methodological approach taken, some challenges arose throughout the course of this project. One notable challenge was reducing inaccuracies in AI responses. At times, the model misinterpreted contextual nuances in client documents or made haste assumptions, resulting in incorrect extractions. Finetuning of prompts and iterative testing was able to refine AI behavior and successfully remediate this recurring issue. Additionally, finding a balance between transparency with efficiency was another large challenge. Detailed step-by-step logic explanations enhanced interpretability but simultaneously raised response times and computational demands.

Striking a balance between clarity and productivity proved essential in optimizing prompt structure. Furthermore, the AI model had to attune to a wide array of document structures. The challenge arose in generalizing the AI's approach across various sources as financial and regulatory documents differ greatly in format and terminology. To ensure consistency and accuracy, prompts were curated to consider common document differences while maintaining uniformity in AI output.

Finally, maintaining data security and compliance was the primary responsibility throughout the project. As the AI system worked with highly sensitive financial information, secure data handling practices and compliance with cybersecurity regulations were critical to mitigate risks associated with unauthorized data access or exposure.

## 4. ANTICIPATED RESULTS

Compliance and security program development will be significantly accelerated by the implementation of AI-driven automation in the due diligence questionnaire (DDQ) lookup process. Dependency on manual lookup procedures and the possibility of human error will diminish by utilizing prompt engineering techniques to improve the accuracy and transparency of the AI model. Due to the length and complexity of the client documents, the compliance teams' previous manual reviewing of files to extract pertinent answers took an average of three hours per document. This time has been lowered to about 30 minutes per document with AI-assisted lookup. As the company serves over 1,000 clients, each with approximately five documents, the estimated total time saved upon full implementation is over 12,500 hours. This equates to a considerable efficiency increase for compliance teams. Consequently, employees can devote more time to higher-order responsibilities like risk assessment and regulatory strategy by automating such tedious tasks, which additionally guarantees more accurate and consistent reporting.

Although the project is still in development and has yet to undergo extensive testing, the current results of the model have already provided significant time savings. Present progress indicates that an estimated 60% of the work has been finished, saving approximately 7,500 hours thus far. These savings will continuously increase as advancements are made, resulting in highly efficient document processing. Furthermore, the AI model's capability to produce detailed justifications for its answers greatly improves transparency in cybersecurity compliance processes and allows firms to excel in satisfying regulatory expectations.

## 5. CONCLUSION

This project demonstrates a significant step toward resolving the black box problem in AI systems used for cybersecurity compliance, specifically in the context of due diligence document analysis. Through implementing structured prompt engineering and explainability layers, the model improved in both accuracy and transparency while decreasing the time and effort required for manual document review. These enhancements foster accountability and trust in AI-assisted workflows, which are critical for high-stakes industries like cybersecurity and finance. The insights gathered from this implementation provide a solid foundation for future XAI advancements and presents new opportunities for responsible AI development centered on ethical integrity, clarity and efficiency.

## 6. FUTURE WORK

Future steps will involve incorporating advanced Explainable AI (XAI) methods such as Shapley (ES) Values and Local Interpretable Model-Agnostic Explanations (LIME) to further improve the AI model's interpretability and transparency. These techniques will provide users with a high depth understanding of the system's decision-making process by measuring the influence of distinct input features on the generated output. To evaluate the explainability framework's scalability and adaptability across various use cases, the system can be examined for extension to additional cybersecurity tasks, such as threat detection, anomaly analysis, and compliance auditing. To ensure model responses maintain accuracy and clarity across varied operational environments, thorough user testing and evaluation will additionally be essential next steps.

## REFERENCES

[1] Bélisle-Pipon, J., Monteferrante, E., Roy, M., & Couture, V. (2023). Artificial intelligence ethics has a black box problem. *AI & Society, 38*(4), 1507-1522. doi:https://doi.org/10.1007/s00146-021-01380-0

[2] Brożek, B., Furman, M., Jakubiec, M. *et al.* The black box problem revisited. Real and imaginary challenges for automated legal decision making. *Artif Intell Law* 32, 427–440 (2024). https://doi-org.proxy1.library.virginia.edu/10.1007/s10506-023-09356-9

[3] Taddeo, M. (2019). Three ethical challenges of applications of artificial

intelligence in cybersecurity. *Minds and machines*, *29*, 187-191.

[4] Al-Mansoori, S., & Salem, M. B. (2023). The role of artificial intelligence and machine learning in shaping the future of cybersecurity: trends, applications, and ethical considerations. *International Journal of Social Analytics*, *8*(9), 1-16.