Enhancing AI Transparency in Cybersecurity: Tackling the Black Box Problem Through Explainable AI Solutions

Unpacking AI's Black Box: An Actor-Network Theory Analysis of Transparency and Accountability in AI Systems

A **Thesis Prospectus Submitted to the**

Faculty of the School of Engineering and Applied Science

University of Virginia • Charlottesville, Virginia

In Partial Fulfillment of the Requirements of the Degree

Bachelor of Science, School of Engineering

Caroline Coughlin

Fall, 2024

On my honor as a University Student, I have neither given nor received unauthorized aid on this assignment as defined by the Honor Guidelines for Thesis-Related Assignments.

ADVISORS

Richard D. Jacques, Department of Engineering and Society

A. Introduction

In our modern world, artificial intelligence has become embedded in many aspects of our lives. With AI's rapid development in the last decade, numerous machine learning-based tools are currently used in daily decision-making processes. These tools are built to process large datasets, identify underlying patterns, and propose solutions. Some common applications of AI-generated solutions include resolving legal disputes, suggesting medical treatments, and approving financial loans. Although these tools simplify major tasks, supplying convenience to our lives, one must simultaneously recognize the widespread implications that exist for using these systems in practice. A significant challenge with these AI-driven solutions is the difficulty in determining how or why the algorithm arrived at the offered resolution. Due to the uncertainty and lack of transparency surrounding algorithms, humans are unable to visualize and interpret how deep learning systems reach their decisions and predictions (Belisle-Pipon et al., 2023). High-stakes fields such as healthcare, finance, or legal areas are most at risk. Errors may result in profound consequences and the reasoning behind judgments has the power to alter outcomes in these fields. My technical project examines these challenges by reflecting on my internship experience with a cybersecurity company, where I contributed to an AI project. My work on this project entailed applying prompt engineering to optimize AI model responses for greater accuracy and reliability. Through this work, I gained firsthand insight into the importance of enhancing transparency and understandability in AI outputs. Building on this understanding, I will examine and explain potential solutions to the black box problem such as improving algorithmic transparency, enhancing legal frameworks, and developing explainability standards.

For my STS project, I will analyze the black box problem in artificial intelligence through the lens of Actor-Network Theory (ANT). Using ANT as the framework, I will investigate how human and non-human actors—such as developers, users, algorithms, and data sources—interact and shape the dynamics of transparency and accountability in AI systems, uncovering where and why opacity persists in these networks.

B. Technical Project

The "black box" problem refers to how AI models take input and produce output with limited or unclear explanations of their decision-making process (Von Eschenbach, 2021). AI scholars inscribe the black box problem under the concept of explainability. If an AI system's functioning can be reduced to a simplified external representation and receives human comprehension, it is said to be explainable (Brozek et al., 2024). Both technical and legal perspectives attempt to address the issue of explainability, however, a gap remains between the two. The law does not explicitly define explainability but offers requirements linked to explainable decision-making, specifically in automated systems. These rules vary across sectors (private vs. public) and contexts (automated vs. general decision-making), making explainability a fragmented concept. Explainability is increasingly regulated for automated decision-making in the private sector, especially by European laws like the GDPR. These regulations require that AI systems reveal key features used in decision-making, but interpretations vary. Some call for revealing all data features used in a decision, while others advocate for disclosing the entire model.

At present, the field of artificial intelligence is rapidly evolving while legal and ethical issues continuously arise, and regulations fail to keep pace. Without proper remediation of the black box problem, significant consequences will ensue such as a lack of accountability for AI-created decisions, errors or biases within data and models, and an erosion of trust. These effects additionally increase the difficulty for widespread AI adoption in all fields. In legal contexts, a lack of explainability can lead to wrongful decisions with severe consequences, such as unfair sentencing. With regulatory laws such as the GDPR, AI systems that cannot provide explanations for automated decisions can face legal challenges or penalties (Brozek et al., 2024). As a result, businesses can also face operational risks for their use of machine learning models.

Throughout my internship experience, I witnessed the black box problem and its effects firsthand. I worked on an AI application that scanned client documents to extract answers for due diligence questionnaires, addressing crucial cybersecurity needs. This AI model often provided inaccurate questionnaire answers with little or vague explanations as to its decision process. The first step of the prompt engineering process was to analyze the locations the AI cited for its answer retrieval along with the accuracy of each questionnaire response in order to identify discrepancies. After examining the documents, I addressed the AI's common mistakes and curated detailed prompts as input to feed the model. These prompts were questions and instructions, which guided the AI to finding accurate and reliable questionnaire answers in the documents. The next set of prompts I developed were designed to target the black box problem and provide explanation behind the AI's responses. I designed specific instructions, detailing every logical step that I wanted the AI to vocalize to the user. Upon adding these prompts to the model's knowledge base, the model provided a chronological visualization of the AI's thought processes behind each decision.

Overall, my work significantly improved the reliability and interpretability of the AI's outputs, reducing the risk of inaccurate information reaching clients—a critical concern in cybersecurity. In this field, incorrect answers in due diligence questionnaires may lead to compliance issues or security vulnerabilities. By tackling the model's opacity, my contributions helped enhance trust in the AI application, ensuring that users could not only rely on accurate answers but also understand the reasoning behind each decision, which is essential in high-stakes fields like cybersecurity. This project inspired me to explore strategies beyond prompt engineering to make AI systems more transparent and accessible to end-users.

Expanding on the goal of increasing AI transparency, various methods have been developed to address the black box problem and enhance the understandability of algorithms. These methods fall under Explainable AI (XAI): a set of techniques that provide clear explanation as to the decision-making process of AI models for human users (Zednik, 2021). One widely used technique is the Local Interpretable Model-Agnostic Explanations (LIME), which works by training local surrogate models to estimate specific model predictions. In essence, LIME creates a new dataset of perturbed samples around an instance of interest, using a weighted proximity metric to train a simpler, interpretable model that simulates the complex model's behavior locally (Hassija et al., 2024). An extension of LIME, known as Anchors, improves the efficiency of LIME by precomputing explanations for a subgroup of instances, which can then be used to develop explanations for the remaining instances in the dataset. Additionally, Shapley (ES) Values connect several XAI techniques, such as LIME and DeepLIFT, and allow us to determine each method's contribution to the prediction. Lastly, model-agnostic approaches operate without reliance on internal model parameters and provide an intuitive, visual representation of the AI's decision logic. Evidently, XAI methods are powerful strategies in helping end-users understand AI thought processes and have made significant progress thus far in improving AI model interpretability.

C. STS Project

My STS research project serves as an extension of my technical project, allowing me to examine the black box problem from a broader, interdisciplinary perspective. While my technical project focused on improving AI transparency in a cybersecurity setting and studying XAI solutions, this research project aims to step back and analyze the origins and implications of the black box issue. Using Actor-Network Theory (ANT) as a framework, I will explore how various human and non-human actors, including technical systems, algorithms, and philosophical viewpoints, interact to shape the challenges and possibilities associated with AI transparency and reliability. This approach will offer insight into the broader social and technical networks that influence AI development and application.

The Actor-Network Theory was designed to comprehend how scientific and technological concepts emerge and gain influence. The primary goal of using this theory is to visualize how people, ideas, technologies, and nature form networks. Relations within the network are considered to be newly formed rather than inherent, arising through ongoing interactions within the network, and needing continuous reinforcement. From an ANT lens, any and all entities in the social and natural world are part of a constantly shifting network of relationships. These entities–people, technology, and objects–are all equally weighted in importance. Therein lies the assumption that there are no external social forces acting on the network and that all components in a social scenario exist on a uniform level. Actors are defined as any entity that influences a techno-social system, including both human and non-human participants. ANT views actors, networks, and systems as interconnected and collaboratively shaped concepts.

While many recognize the value in perceiving technical concepts through ANT, others place critiques on the framework. Critics of ANT argue that it overly emphasizes non-human entities as having agency and reduces people to simple positions in networks without real influence. They also assert that ANT is overly descriptive, ignoring social factors like race and class, and relies on subjective choices to determine actor relevance, which can lead to endless associations without clear explanations.

From an ANT perspective, the black box problem in AI can be understood as a network of interacting elements—humans, algorithms, data, infrastructure, and regulatory frameworks—whose relationships and roles are continuously reshaped as AI systems evolve. The black box phenomenon itself is not just a technical issue but one deeply embedded in a socio-technical network. In this view, AI systems are not isolated entities that autonomously "act" on their own; rather, they emerge and gain influence through their interactions with other actors, such as developers, corporations, policymakers, and the public. These interactions define the opacity of AI systems, as each actor brings its own interests, assumptions, and limitations to the network, which contributes to the complexities surrounding AI transparency.

One example of this can be seen in the way algorithms and data are constructed and employed. While AI systems are designed and coded by human engineers, the training data used to feed these systems often contain biases or represent only a narrow range of human experience. These biases are then propagated and magnified by the algorithms, leading to outcomes that may be inaccurate, discriminatory, or unjust. However, these biases are not solely the responsibility of the AI systems themselves but rather emerge from the decisions made by human actors in the design, development, and deployment phases. The black box, in this case, is a product of the networked actions and relationships of those involved, including the stakeholders in the data collection and model training process, the ethical frameworks they adhere to (or neglect), and the regulatory bodies that may or may not intervene to ensure accountability.

Another key component of the black box problem is the power dynamics within the networks of AI development. For example, major tech companies that build AI systems often hold a disproportionate amount of power in shaping the discourse around AI transparency and accountability. These companies can control access to the algorithms they develop, citing proprietary concerns or security risks, further obscuring the internal workings of these systems. From an ANT perspective, this control is a key part of the network of power that shapes the black box problem. While AI developers, policymakers, and users are aware of the importance of transparency, their ability to address it is often hindered by the larger network of interests that prioritize profit and competitive advantage.

The significance of this research lies in its potential to illuminate the social forces and human decisions that contribute to the opacity of AI systems. By examining the actors involved in the development and deployment of AI technologies through an ANT lens, this project seeks to highlight the ways in which human and non-human entities contribute to the existence of the black box problem. It emphasizes the need for more nuanced, ethnographic approaches to understanding algorithmic systems, focusing not only on the technology but on the broader socio-technical context in which it exists (Christin, 2020). Understanding the black box from this perspective can lead to more informed conversations around AI ethics, policy, and regulation, promoting a more transparent and accountable approach to AI development. As technology becomes increasingly integrated into societal systems, recognizing and addressing the black box problem through ANT could offer valuable insights into how we can mitigate the risks and maximize the benefits of AI.

D. Conclusion

As I am a computer science student, there is no technical deliverable.

The final objective of my STS research is to build an argument that examines the black box problem in AI through the Actor-Network Theory (ANT) framework, focusing on how socio-technical dynamics influence AI transparency and accountability. To support this argument, I will use two case studies that demonstrate the intricate connections between human and non-human actors in AI systems, highlighting how relationships among developers, data, algorithms, and policy influence the development and deployment of these technologies.

My goal for this STS research project is to provide a nuanced understanding of AI opacity and to advocate for a more accountable and transparent approach to AI development. By framing AI within an ANT lens, I aim to contribute to discussions on AI ethics, policy, and regulation, emphasizing the importance of considering both technological and human factors in addressing AI's black box problem.

2179 words

References

Adadi, A., & Berrada, M. (2020). Explainable AI for healthcare: from black box to interpretable models. In *Embedded systems and artificial intelligence: proceedings of ESAI 2019, Fez, Morocco* (pp. 327-337). Springer Singapore.

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

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

Carabantes, M. (2020). Black-box artificial intelligence: an epistemological and critical analysis. *AI & society*, *35*(2), 309-317.

Christin, A. (2020). The ethnographer and the algorithm: beyond the black box. *Theory and Society*, *49*(5), 897-918.

Durán, J. M., & Jongsma, K. R. (2021). Who is afraid of black box algorithms? On the epistemological and ethical basis of trust in medical AI. *Journal of Medical Ethics*, *47*(5), 329-335.

Hassija, V., Chamola, V., Mahapatra, A., Singal, A., Goel, D., Huang, K., ... & Hussain, A. (2024). Interpreting black-box models: a review on explainable artificial intelligence. *Cognitive Computation*, *16*(1), 45-74.

Khrais, L. T. (2020). Role of artificial intelligence in shaping consumer demand in E-commerce. *Future Internet*, *12*(12), 226.

Quinn, T. P., Jacobs, S., Senadeera, M., Le, V., & Coghlan, S. (2022). The three ghosts of medical AI: Can the black-box present deliver?. *Artificial intelligence in medicine*, *124*, 102158.

Rai, A. (2020). Explainable AI: From black box to glass box. *Journal of the Academy of Marketing Science*, *48*, 137-141.

Sharma, S., Henderson, J., & Ghosh, J. (2020, February). CERTIFAI: A common framework to provide explanations and analyse the fairness and robustness of black-box models. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* (pp. 166-172).

Von Eschenbach, W. J. (2021). Transparency and the black box problem: Why we do not trust AI. *Philosophy & Technology*, *34*(4), 1607-1622.

Wadden, J. J. (2022). Defining the undefinable: the black box problem in healthcare artificial intelligence. *Journal of Medical Ethics*, *48*(10), 764-768.

Wang, F., Kaushal, R., & Khullar, D. (2020). Should health care demand interpretable artificial intelligence or accept "black box" medicine?. *Annals of internal medicine*, *172*(1), 59-60.

Yang, G., Ye, Q., & Xia, J. (2022). Unbox the black-box for the medical explainable AI via multi-modal and multi-centre data fusion: A mini-review, two showcases and beyond. *Information Fusion*, *77*, 29-52.

Zednik, C. (2021). Solving the black box problem: A normative framework for explainable artificial intelligence. *Philosophy & technology*, *34*(2), 265-288.