Published by :
https://www.ijert.org/
An International Peer-Reviewed Journal

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14 Issue 12 , December - 2025

The Black Box Opacity and Theoretical Understanding of Generative Ai Setbacks

Dr. Shahebaz Ahmed Khan
Associate Professor, Jayaprakash Narayana Engineering College, Hyderabad

Dr. Mohd Abdul Qadeer
Assistant Professor, Avanthi Institute of Engineering and Technology, Hyderabad

Dr. Abdul Ahad Afroz
Assistant Professor, ISL Engineering College, Hyderabad

ABSTRACT - This paper attempts to analyze and examine the observed deficiencies in the field and technology of Artificial Intelligence. Despite some extensive potential, the systems of Artificial Intelligence face significant limitations in terms of various constraints of technology, ethics and generalization. The AI has such limitation due to information processing complexity and its diverse conjunctional elements. Also, the selection pressures relating to tradeoffs can put forth the ethical, structural and technical drawbacks associated with AI technology. In AI, the black box opacity erodes trust and accountability by limiting the performance besides the other issues like bias, dependency and generalization. In AI systems, particularly the deep learning models without offering any clear or proper insights about decision making and conclusions remains a mystery and paradox to its users. Though Artificial Intelligence is a defining technology, considered as the driving force of digital revolution, there is a landscape limitations and obstacles which hinder the implementation and practical advancements of AI systems. Before the complete potential of AI is analyzed and realized, this paper undermines unresolved limitations requiring cross-disciplinary innovation. The adoption and integration of Artificial Intelligence in real world environments can also create unnecessary and unwanted outcomes. The black box opacity poses a challenge leading to some unfair and discriminatory outcomes in the context of addressing potential biases whenever the decision-making process is not transparent, but opaque. It can also reduce the trust in model outputs causing security and ethical concerns with difficulty in operations.
Key words: Black box opacity, generalization, potential bias, information processing, defining technology etc.

1. INTRODUCTION

The system of AI is capable of potential automation and finite analysis, but at the same time it is limited to its functions constrained by low contextual understanding, algorithmic biases, data quality, and interpretation of data transparency during decision making. Today we find, powerful generative AI models which primarily rely upon complex neural networks intended to develop response for the processing of natural language. Here, the question is interpreting the internal functioning of those networks and knowing what exactly is happening inside the model. The rule-based AI models when deployed can be easy to understand, but may not be so powerful and flexible in terms of complex function implementation in comparison with generative AI models [1]. In such a way, the deep neural networks in nature are inherently opaque. These complex black boxes in advanced AI systems can deliver impressive results, but they lack transparency and accountability where it becomes hard to trust their outputs.

The black box opacity of AI systems forces us to think over the issues loke accountability, performance, explain ability, ethical concerns and security [2]. Not only these parameters make humans to reconsider the use of AI systems in real world setting, but also the environmental cost as well as financial scales can make unsustainable deployments. Intricately, the reliability, performance and accountability of AI is combinatorial to the data used to train the models which we deploy in AI systems. This primary and core dependency on huge data lacks the generalizability and keeps improper bias causing significant failures with wrong outcomes further making difficult to operate the environment and design implementation. The users of Artificial Intelligence must be aware of its inadvertent negative ripples and backlashes restricting the human behavior, identity and autonomy. Despite unimaginable advancements and abilities to define real world environments, still the Artificial Intelligence systems suffer various negative factors such as data bias, lack of interpretability, regulatory noncompliance, reduced trust etc. and pose significant challenges like security concerns, ethical concerns and accountability. The opacity in some complex AI systems make it difficult to understand the arriving of conclusions by a system, error generation verification and broad interpretation of operations lacking transparency. The lack of transparency can be a significant challenge during automation in industries and other fields where real time decision making is only option. There may be some practical reasons and validations to use black box models of deep learning, but the consequences of black box opacity invite severe practical and ethical vulnerabilities preventing

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14 Issue 12 , December - 2025

effective auditing and potential systemic risks at the expense of infrastructure costs, environmental damage and computational power.

2. THE OPACITY CONUNDRUM

A vexing challenge and inability in generative AI is to understand the internal logic, reasoning and functional aspects of complex machine learning models used in it. This means the details are in black box. This nature of black box undermines the accountability, crates significant security risks and complicates the problem of debugging.

2.1 The Black Box Quandary

The black box opacity in Artificial Intelligence refers to dearth and absence of transparency, where the internal decision-making processes are hidden and are not understandable to the users. In black box opacity, we can find the inputs and outputs, but the algorithmic logic, reasoning and internal functioning theme is absent to our understanding. Many machine learning models like ChatGPT, Open AI, DeepSeek, Meta AI etc. are based on black box AI. These models are trained on huge amounts of data to undergo deep learning processes. Sometimes, even the makers and creators are unaware of the complete picture of model processing and internal working. The black box can give impressive outcomes, but one cannot see how internal results, decisions, classifications, predictions are made. The black box models involve millions of parameters, advanced algorithmic logics and processing of many deep layers. Due to such opaque nature of models, the users cannot accurately validate the outcomes of the models being unaware of internal logical functioning. Further the black box models of deep learning can also hide vulnerabilities, can create bias issues and violate the privacy in certain aspects of AI system deliveries. So, it makes the users to trust the outcomes hardly. The working of hidden layers remains as a mystery in many situations and gradually the powerful AI system lacks accountability and interpretability. Gradually, the tradeoff problem arises between the performance and transparency. The traditional AI models make use of more interpretable models like regression and tree-based models but may not be accurate with prediction and lack validation.

The black box opaque is common in models of deep neural networks [3] [4]. The deep neural networks take unstructured, raw and massive data for analysis. These networks are capable of identifying the patterns and learning from these identified patterns for future use. There is no supervision but a little human intervention during training of an AI model which makes these models capable of processing languages, creating original content and acquiring human intelligence to an extent. The developers of Artificial Intelligence can come to know how the data is processed through each layer in the network, but are not found with some specific logical constraints that enable the AI systems to gain such impressive abilities of decision making. For example, the creators of AI may not understand how exactly a model defines, finds and combines the vector embedding to a prompt for the response. Figure 2.1 shows the basic model of black box AI dilemma.

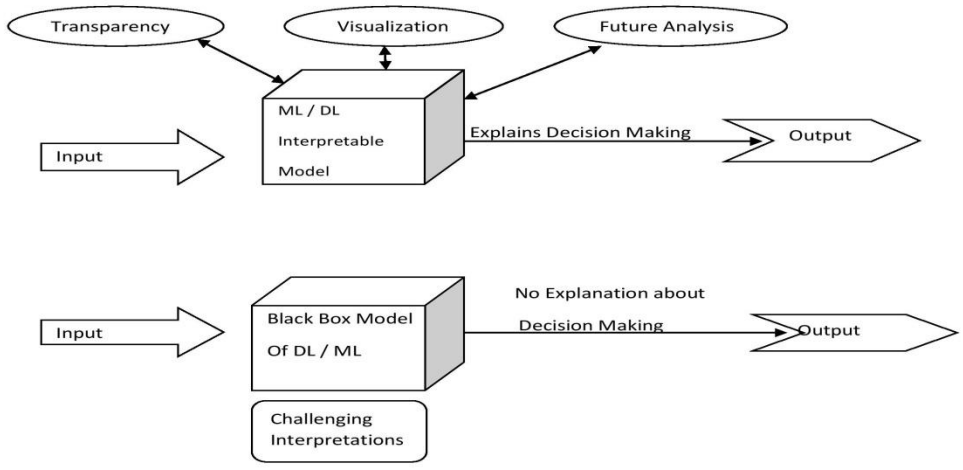

Figure : 2.1 Black Box Dilemma

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14 Issue 12 , December - 2025

2.2 Reasons for Black Box AI Models

Primarily, there are two reasons to find black box AI systems [5]. One is that, with a specific purpose the developers make a model into a black box and the second, the models take the shape of black boxes as a by product from or after the training. To protect the intellectual property, some developers recondite the internal working of the AI model, they know how the model works but intentionally tend to keep the design, source code and decision making as a secret. In some cases like generative AI systems, the creators intentionally do not hide the operation of internal structure, but the processing of deep networks make these models so powerful and complex where even the developers or creators fail to understand what exactly happens inside the hidden layers. These are sometimes called organic black boxes.

2.2 Opacity Consequences and Manifestation

The black box opacity can be correlated to the sight of hallucinations. We know that the AI models are meant to receive continuous input sequences and perform training. Whenever a trained model on some generic data is applied to a particular data set, then we can outcomes can be inaccurate and performance can be unreliable. Due to this scenario, we can find the emerging of biases and errors termed as hallucinations [6]. This training can be indefinite sometimes and the data ingesting is formally not visible and can start doing some unexpected actions on the data that even confuse the creators which can be termed as hallucinations.

A survey has shown that the factors and parameters which contribute the black box phenomena and interpretability have become a primary reason for the failure of AI projects in various organizations. The severe practical and ethical consequences of the black box are the mitigation of trust in validating results. The AI outcomes cannot be taken into confidence due to lack of transparency, this transparency is under scored as the users cannot validate the model upon the condition of not knowing what is happening with the data internally. If the results are bounded in confidence boundary, then it can potentially cause the altogether failure of AI implementation in a particular system. Today, the liability and accountability vacuum in implementation of some complex and advanced AI projects is questionable due to the opacity of black box phenomenon in AI systems. There have been and can be severe ethical, environmental and cyber security consequences in this context, also privacy violations pose a big challenge in use of such AI systems. The limited visibility and understanding of AI operations do not provide the clear insights about the system behavior and agent steps further which threatens the data security and integrity used in that environment.

The models found with black box opacity when modified at error inputs can significantly impact the generated output quality in negative terms making it more vulnerable to cyber attacks. The researchers and AI experts are actively finding the ways to solve the problems of black box using ‘intrinsic interpretability’ where models that are inherently transparent in decision-making process are easy to understand and ‘post hoc techniques’. The post hoc techniques like LIME **(**Local Interpretable Model-Agnostic Explanations)**,** SHAP **(**Shapley Additive Explanations), Guided GradCAM and PDPs (Partial Dependence Plots) can be local that can explain individual predictions and can also be global which can explain the complete model. These techniques can be useful in building the trust in model validations, identifying biases, debugging etc in some subtle and sensitive areas like medical care, health industry, banking and finance [7]. Here, improving descriptive accuracy interpretability is a challenging criterion which can be achieved at the expense of predictive accuracy which further can be gained only through model performance. This often leads to a trade-off in AI system design and implementation costs.

3. TECHNICAL PERFORMANCE AND OTHER CONSTRAINTS

Besides the data and cost limitations, the current AI systems are found with innate technical barriers which clampdowns the reliability, ability and capacity. The black box opacity paves a way to challenges related to generalization, contextual understanding and system degradation. AI systems are often designed to perform specific tasks and lack the ability to generalize beyond their pre-programmed functions. The black box theme is responsible for erosion of human connections and loss of human autonomy [8]. Complex AI systems which involve black box opacity are in need of large amounts of structured or standardized data which comes with the human right to privacy in some cases. One should also note a point that, the AI that operates on intrinsic interpretability is not of general human intelligence, but a powerful pattern matching and extrapolation machine. The AI with human intelligence replacement primarily considers the idea of black box opacity though unintentionally which further derives the setbacks as discussed above.

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14 Issue 12 , December - 2025

4. DEALING WITH BLACK BOX OPACITY

Organizations can deploy transparent models wherever possible, but this cannot be possible in all workflows and cases. Some workflows require complex black box AI models. But there are some ways to mitigate the black box risks and make black box models more trustworthy.We can use open-source models which are more transparent when used for development and operations. At the same time, an open-source generative AI model might be a black box due to its complex neural network [9], but it can offer users more insight than a closed-source model.

AI governance can also be helpful to avoid black box to some extent, the AI governance enables the AI creators to establish strong control structures by limiting the complex operations with powerful outcomes. The AI governance tools monitor and audits the trails to detect anomalies. But, this may not be applicable to all cases of AI systems. Responsible AI frame work also can deal with black box opacity. It provides a set of principles to an organization to deploy a model in transparent way upon practicing can make AI more trustworthy.

5. CONCLUSION

The problem of black box opacity in generative AI where the internal operations remain in opaque creates a dilemma of accountability and trust. For Artificial Intelligence, the black box opacity limitations may not be primarily computational but are deeply related to data, logic, finance, environment and ethics. The black box situations can cause unpredictable failures in the AI systems. The paper theoretically concludes and suggests that the current Artificial Intelligence is merely a powerful tool for deep pattern matching and automation, but not a just or absolute form of human intelligence. So, the AI in further developments requires a careful monitoring, management and deployment. The challenges posed by black box AI systems require a delicate balance between the renovation and transparency by emphasizing the importance of openness in AI development.

6. REFERENCES

[1] https://jolt.law.harvard.edu/assets/articlePDFs/v31/The-Artificial-Intelligence-Black-Box-and-the-Failure-of-Intent-and-Causation-Yavar-Bathaee.pdf, Harvard Journal of Law and Technology.
[2] Adadi A, Berrada M. Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). IEEE Access. 2018;6:52138–60.
[3] Tjoa E, Guan C. A survey on explainable artificial intelligence (xai): Toward medical xai. IEEE Trans Neural Netw Learn Sys. 2021;32(11):4793–813.
[4] Raghu M, Gilmer J, Yosinski J, Sohl-Dickstein J. Svcca: Singular vector canonical correlation analysis for deep learning dynamics and interpretability. 2017. https://arxiv.org/abs/1706.05806.
[5] Guidotti R, Monreale A, Turini F, Pedreschi D, Giannotti F. A survey of methods for explaining black box models. CoRR. 2018;abs/1802.01933. http://arxiv.org/abs/1802.01933.
[6] Adhikari A, Tax DMJ, Satta R, Faeth M. Leafage: Examplebased and feature importance-based explanations for black-box ml models. In: 2019 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE). 2019. p. 1–7.
[7] Shrikumar A, Greenside P, Shcherbina A, Kundaje A. Not just a black box: Learning important features through propagating activation differences. CoRR. 2016;abs/1605.01713. http://arxiv.org/abs/1605.0171.
[8] Apley DW, Zhu J. Visualizing the effects of predictor variables in black box supervised learning models. 2019.
[9] Han S, Ding H, Zhao S, Ren S, Wang Z, Lin J, Zhou S. Practical and robust federated learning with highly scalable regression training. IEEE Trans Neural Netw Learn Syst. 2023.

(This work is licensed under a Creative Commons Attribution 4.0 International License.)