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

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 05 , May - 2026

The “Blackbox” Problem in AI

Investigating The Transparency of Deep Learning Algorithms In High-Stakes Decision Making (Like Health And Law)

1.Ms. Joytria Fernandes
Student
B-Tech CE (computer engineering)
GHRISTU, Shastri Nagar, Pune

2.Mr. Parish Jadhav
Student
B-Tech CE (computer engineering)
GHRISTU, Shastri Nagar, Pune

3.Mr. Chirag Talekar
Student
B-Tech CE (computer engineering)
GHRISTU, Shastri Nagar, Pune

4.Mr. Revanth Sai Pagoti
Student
B-Tech AIDS (artificial data sci.)
GHRISTU, Shastri Nagar, Pune

INTRODUCTION

Artificial Intelligence (AI) has become a crucial technology in today’s world. It is used in many fields, including healthcare, law, banking, education, and transportation. AI systems can quickly analyze large amounts of data and make decisions that assist humans. As a result, AI improves efficiency and saves time in various industries.

However, not all AI systems are easy to understand. Many modern AI models, especially deep learning models, operate in ways that are not clear to humans. These models take input data, process it through several layers, and produce an output. But the steps in between remain unclear. This is called the “Black Box” problem.

The black box problem means we can see an AI system’s input and output, but we do not understand how it reached that decision. This lack of clarity leads to significant issues, especially in areas where decisions carry great weight. For example, if an AI system predicts a patient has a disease, doctors need to know how that prediction was made. Without an explanation, it becomes hard to trust the system.

In high-stakes fields like healthcare and law, incorrect decisions can have severe consequences, such as loss of life or unfair judgments. Therefore, understanding and resolving the black box problem is very important.

This report examines the black box problem in AI, its impact on various sectors, ethical issues, and problem in AI, its impact on various sectors, ethical issues, and potential solutions like Explainable AI (XAI).

ABSTRACT

Artificial intelligence has progressed rapidly, but many deep learning systems are still difficult to interpret, leading to the “black box” problem in high-stakes areas such as healthcare and criminal justice. These models can achieve strong prediction performance but their layered, non-linear structure often makes transparency, trust and accountability difficult. The paper examines the conceptual basis of black-box behavior in deep learning and its implications, through the case studies of IBM Watson and COMPAS. It also reviews explainable AI methods like LIME, SHAP as partial solutions for improving model interpretability. Effective deployment of AI in critical decision making requires not only accuracy but also explainability, fairness and human oversight, the study argues.

METHODOLOGY

The study uses qualitative review of academic literature, legal analyses, and investigative reports on the black-box problem in artificial intelligence, explainable AI, fairness, and accountability. Sources were chosen based on their relevance to high-stakes decision making and explicit discussion of transparency-related issues in real-world systems. We chose IBM Watson and COMPAS as case studies. These are two frequently cited examples in healthcare and criminal justice, respectively, that illustrate the impact of opaque AI systems on trust, fairness, and human oversight in high stakes domains.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 05 , May - 2026

To assess these systems, the paper draws on the established framework of trustworthy AI based on transparency, defined as "the communication of the logic and data that underpinned decisions in a comprehensible manner"; fairness, relating to whether the outcomes produced show systematic drawbacks for particular groups; accountability, concerning who is responsible when there is negligence; and interpretability, which means that users and affected parties can understand individual predictions. This methodology involves collecting and analysing published audits, investigative reports, legal rulings, and critiques to give context in understanding real-world impacts rather than re-running models for each system. This evidence-based approach aims to highlight relevant patterns and lessons that apply across cases rather than within one system.

BACKGROUND AND GROWTH OF AI

Artificial Intelligence has changed a lot over the years. In the beginning, AI systems followed simple rules. These systems had clear instructions, making them easy to understand. For example, a rule-based system might state, "If the temperature is high, then turn on the fan." Such systems were straightforward and predictable.

Later, machine learning (ML) came into play. Instead of following fixed rules, ML systems learn patterns from data. For instance, a machine learning model can detect spam emails by analysing previous examples. This advancement made AI stronger but also less clear.

The biggest change happened with deep learning. Deep learning uses neural networks with multiple layers to process complex data like images, speech, and text. These models can achieve high accuracy but are also very complex.

This complexity makes it challenging to understand how these models make decisions. Even the developers of these models may not fully understand their workings. This is where the black box problem becomes important.

Today, AI is used in:

• Medical diagnosis

• Self-driving cars

• Financial predictions

• Legal decision-making

As AI continues to evolve, the demand for transparency grows. Without understanding how AI works, it becomes risky to rely on it in critical situations.

DEEP LEARNING AS A BLACK BOX: THE ROLE OF HIGH-DIMENSIONAL FEATURE SPACES AND NON-LINEAR TRANSFORMATIONS

Deep learning is often viewed as a "black box." Although it predicts with high accuracy, it is difficult for people to understand how inputs connect to outputs. This confusion arises from two related problems: the high-dimensional nature of feature spaces and the stacked, non-linear transformations in hidden layers. These factors together make it hard to interpret individual neuron activations.

High-Dimensional Feature Spaces

Modern neural networks process inputs with thousands or millions of features, such as pixels in images, embeddings in natural language tasks, or high-resolution sensor signals. These inputs push the network into high-dimensional spaces where standard geometric ideas do not work. Distances become almost uniform, most points are nearly orthogonal, and neighbourhoods behave in surprising ways. Within the network, hidden layers learn lower-dimensional, task-specific embeddings, but these embeddings do not necessarily have clear meanings. They simply represent the configurations that minimize the loss function. Because of this, the features the network learns blend many original dimensions, and there is no assurance that any single neuron corresponds to a clear concept.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 05 , May - 2026

Non-Linear Transformations in Hidden Layers

Hidden layers perform repeated non-linear transformations, such as ReLU, sigmoid, or other activation functions, along with affine maps like weight matrices and biases. This process enables the network to model complex, non-linear decision boundaries by combining many simple operations. Each layer modifies the input in a non-linear way, making the overall function from raw input to final prediction a deeply nested, non-linear composition. This complexity prevents a straightforward breakdown into understandable steps. Since the activation of each neuron relies on numerous incoming units that pass through these non-linear functions, its value reflects nuanced, context-dependent interactions rather than a single concept. For example, a neuron in an image classifier may capture a mix of edges, textures, and colors. This helps with predictions but does not easily translate to a recognizable visual feature.

Why Individual Neurons Are Uninterpretable

- In deep networks, neurons typically work with distributed representations. Information relevant to predictions spreads across many neurons instead of being concentrated in one. This means that the activity of a single neuron only carries meaning within the context of the entire group of activations. It is not possible to assign a clear semantic label as one might for manually crafted features. Furthermore, training that relies on gradients aims at enhancing performance rather than making interpretations clearer. As a result, the function of any neuron in a specific prediction is uncertain. At best, techniques like saliency maps, SHAP values, or surrogate models can provide rough explanations

- but these are still just approximations. In summary, deep learning operates as a black box because high-dimensional input spaces and deeply stacked, non-linear hidden layers produce representations that work well for predictions but are too intricate and mixed for easy human understanding.

Christoph Molnar, Interpretable Machine Learning: A Guide for Making Black Box Models Explainable (2019, 2020). This book addresses the lack of transparency in modern machine learning models, including deep neural networks. It presents systematic methods, such as LIME, SHAP, and other model-agnostic techniques, to explain black-box predictions. It is one of the most frequently referenced works in research papers discussing the black-box nature of AI and the interpretability of deep learning.

UNDERSTANDING THE BLACK BOX PROBLEM

The black box problem refers to the lack of transparency in AI systems. Simply put, we cannot see or understand how an AI model makes its decisions.

In deep learning models, data travels through many hidden layers. Each layer carries out complex mathematical operations that are hard for people to grasp. As a result, we only see the final output without knowing the reasoning behind it.

For example, consider an AI system that detects cancer from medical images. The system might indicate that a patient has cancer, but it may not show which part of the image led to that conclusion. This creates confusion and weakens trust.

The black box problem has the following features:

• Lack of explanation

• Hidden decision-making process

• Difficulty in troubleshooting errors

• Low trust from users

This problem becomes more serious when AI is used in high-risk areas. If an AI system makes a wrong decision, it can be hard to determine why it happened. This complicates the process of fixing the system.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 05 , May - 2026

IMPACT IN HEALTHCARE

AI is widely used in healthcare for tasks like disease diagnosis, medical imaging, and treatment recommendations. These systems can quickly analyse patient data and offer accurate predictions.

However, the black box problem poses significant challenges in healthcare.

Doctors need to understand the reasoning behind any medical decision. If an AI system suggests a treatment but doesn't explain it, doctors may lose trust. This can lead to rejecting helpful technology.

Another concern is patient safety. If an AI system makes an incorrect prediction, it can result in inappropriate treatment. Without knowing the cause, fixing the issue is challenging.

For example, IBM Watson was used for cancer treatment recommendations. Doctors found that the system sometimes gave incorrect suggestions and lacked clear explanations for its decisions. As a result, trust in the system declined.

This shows that transparency is essential in healthcare. AI should assist doctors rather than replace their judgment without clear reasoning.

CASE STUDY REFLECTION

Looking back at IBM Watson and COMPAS, two things stand out: both systems aimed to help but also exposed the deeper human costs of the black box problem in AI. IBM Watson promised to help doctors make more informed decisions, especially in complex fields like oncology. However, many clinicians found it hard to understand how the system came to its recommendations, making it hard to trust or challenge its outputs. When Watson's suggestions were wrong or seemed random, the lack of clear reasoning put doctors in a tough spot: follow an unclear recommendation, ignore it, or override it without fully understanding the consequences.

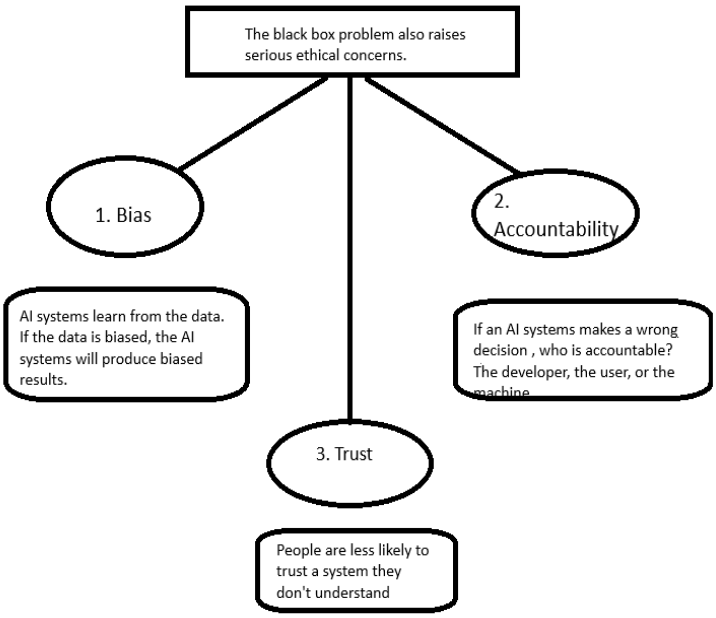
COMPAS tells a different but related story. In the criminal justice system, risk assessment scores were meant to support fairer decisions about bail, sentencing, and parole. However, because the model was opaque, people affected by it—defendants, families, and even judges—could not understand why a high-risk score had been assigned. This lack of clarity created feelings of powerlessness and unfairness, especially when audits later revealed bias against certain groups. The problem was not just technical; it was deeply human. People were being judged by algorithms they could not see, understand, or challenge.

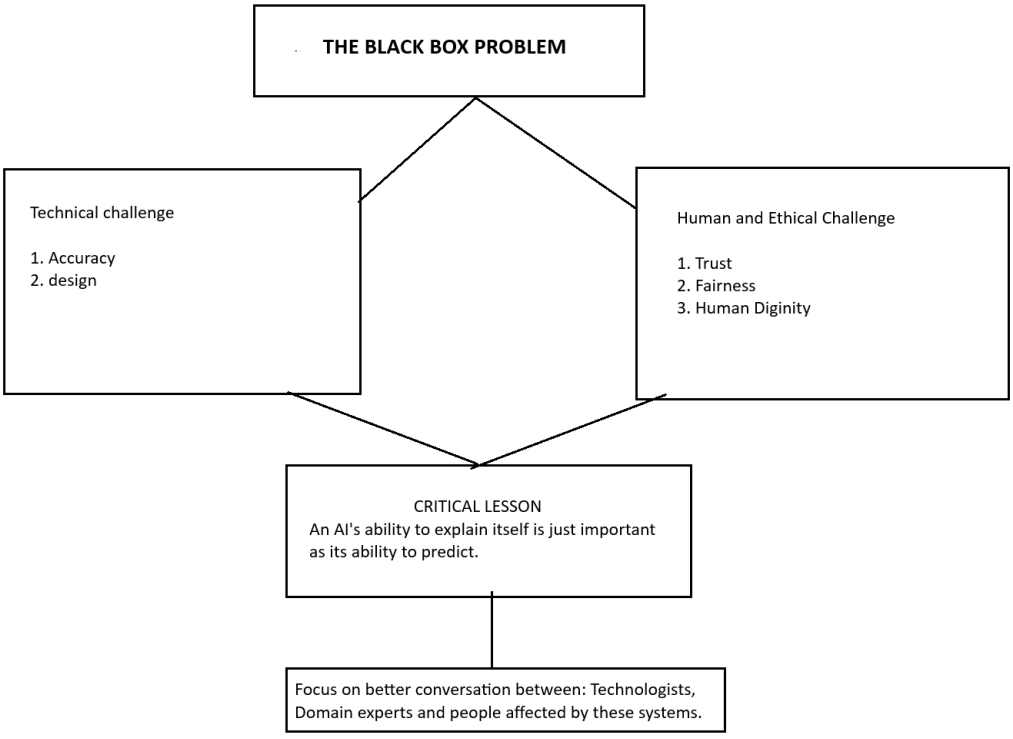
Together, these cases remind us that the black box problem is not just about accuracy or design; it's about trust, fairness, and human dignity. They show that in critical areas, a system's ability to explain itself is just as important as its ability to predict. Moving forward, any efforts to address the black box nature of AI must focus on better conversations between technologists, domain experts, and the people affected by these systems.

DIMENSIONS	IBM WATSON	COMPASS
Impact	Human oversight is essential	Due and equity concerns
Lesson	Reduce patient harm	Transparency are essential
Domain	Healthcare	Criminal justice
Main purpose	Treatment support	Risk assessment
Main concern	Incorrect recommendations	Biased concerns
Transparency issue	Unclear logic	Scoring logic

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 05 , May - 2026

ETHICAL ISSUES IN AI





EXPLAINABLE AI (XAI) – SOLUTION

Explainable AI (XAI) provides a solution to the black box problem. It strives to make AI systems more transparent and understandable.

XAI techniques help explain:

• Why a decision was made

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 05 , May - 2026

• Which factors were significant

• How different inputs affect the output

Two popular techniques are:

• LIME

• SHAP

These methods offer insights into how AI models work. This builds trust and supports better decision-making by users.

CHALLENGES IN EXPLAINABILITY

Although XAI is beneficial, challenges remain.

The biggest challenge is the trade-off between accuracy and clarity. Highly accurate models tend to be complex and hard to explain. Simpler models are easier to understand but may not perform as well.

Other challenges include:

• Complexity of deep learning models

• Lack of standardized methods

• Difficulty explaining real-time decisions

Despite these challenges, research in XAI is advancing rapidly.

CONCLUSION

The black box problem is a major challenge in modern AI systems. While AI offers powerful solutions, its lack of transparency limits its use in critical sectors.

In areas like healthcare and law, decisions must be clear, fair, and reliable. Without explanations, AI systems can pose more risks than benefits.

Explainable AI presents a promising solution by making AI systems more transparent. However, further research is necessary to improve these methods and make them practical.

In the future, AI systems should be designed with both accuracy and clarity in mind. Only then can we fully trust and effectively use AI in important decision-making processes.

REFERENCES

.

[1] Berkeley, L. (2017). State of Wisconsin v. Loomis: Algorithmic risk assessment and the due-process concerns raised by COMPAS. Harvard Law Review, 130(8), 2117–2145.
[2] Kusner, M. J., Loftus, J., Russell, C., & Silva, R. (2017). Counterfactual fairness. Advances in Neural Information Processing Systems (NeurIPS), 30.
[3] Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2018). A survey of methods for explaining black box models. ACM Computing Surveys, 51(5), Article 93.
[4] Hassija, V., et al. (2024). Interpreting black-box models: A review on explainable artificial intelligence. Cognitive Computation, 16(1), 45–74.
[5] Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). IEEE Access, 6, 52138–52160.
[6] Holzinger, A. (2016). Interactive machine learning for health informatics: When do we need the human-in-the-loop? Brain Informatics, 3(2), 119–131.
[7] Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., & Floridi, L. (2016). The ethics of algorithms: Mapping the debate. Big Data & Society, 3(2).
[8] Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. Nature Machine Intelligence, 1(9), 389–399.
[9] Barocas, S., Hardt, M., & Narayanan, A. (2019). Fairness and Machine Learning. Fairness-in-ML.org.
[10] Dieterich, W., Mendoza, C., & Brennan, T. (2016). COMPAS risk scales: Demonstrating accuracy equity and predictive parity. Northpointe Inc. (ProPublica analysis).
[11] Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016). Machine bias: There's software use.