iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 14 **Issue:** III **Month of publication:** March 2026

DOI: https://doi.org/10.22214/ijraset.2026.78828

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue III Mar 2026- Available at www.ijraset.com

Transparency and Explainability in Responsible AI: Foundations, Challenges, and the Path Forward

Siddhika Sachdeva[1], Yogita Thareja[2]
Vivekananda Institute of Professional Studies – Technical Campus

Abstract: ***AI systems now make or heavily influence decisions about who gets a loan, who is flagged as a flight risk, and which patients receive certain treatments etc. Given these stakes, one question keeps coming up in both policy and engineering circles: do we actually understand how these systems reach their conclusions? This paper focuses on two related ideas that sit at the heart of responsible AI: transparency, meaning how open a system is about its inner workings, and explainability, meaning how well it can articulate its reasoning to the people affected by it. I survey the main technical approaches, examine why they fall short in practice, and argue that solving this problem requires more than better algorithms,it requires rethinking how AI systems are governed.***

I. INTRODUCTION

Not long ago, the idea of an algorithm deciding whether someone qualified for parole or a mortgage would have seemed like science fiction. Today it is routine. These systems process far more information than any human reviewer could, and in many settings, they perform impressively well. But there is a growing unease shared by researchers, regulators, and the public — about what happens inside the black box.

The concern is not simply academic. When a hiring algorithm screens out a qualified candidate, or when a predictive policing tool steers resources toward a neighbourhood based on patterns no one can fully explain, the question of accountability becomes urgent. If nobody can explain why a decision was made, it becomes very difficult to challenge it, correct it, or prevent it from happening again.

Transparency and explainability are the main responses the field has developed to this problem. The two terms are related but distinct. Transparency refers to how open a system is — whether its training data, architecture, and parameters are accessible. Explainability is narrower: it asks whether we can articulate, in terms a non-expert can understand, why a specific output was produced. Both matter, and the research on each has grown substantially over the past decade.It is worth being precise about what these words mean, because they are used loosely in both academic and industry contexts.

A cardiologist reviewing an AI-assisted diagnosis needs to know whether the model weighted the right clinical indicators. A loan applicant denied credit wants to know what they could do differently. A regulator auditing a hiring system wants to know whether the model's behaviour correlates with protected characteristics at the population level. The same technical system has to serve all three explanatory needs, and they do not always point in the same direction.

The most direct response to opacity is to use a model that is not opaque in the first place. Linear regression, decision trees, and rule-based classifiers are easy to inspect because their logic is explicit, you can read off the feature weights or follow a decision path and understand exactly what drove the output.

There is a real tension here. Simpler models are sometimes genuinely less accurate, particularly in domains with complex, high-dimensional data. But the accuracy gap is often smaller than practitioners assume, and in high-stakes settings the value of legibility may outweigh small performance differences. It is worth asking more often whether interpretability-first is an option before reaching for a more complex architecture.

When a complex model is already in place, post-hoc methods try to explain its outputs after the fact. Two have become especially prominent. LIME (Ribeiro et al., 2016) works by perturbing the input around a specific prediction and fitting a simple model to the resulting outputs, producing a local approximation of the black-box model's behaviour in that region. SHAP (Lundberg & Lee, 2017) draws on the Shapley values from cooperative game theory to assign each input feature a contribution score that reflects how much it pushed the prediction in a given direction.

Both methods are widely used and genuinely useful. But they come with caveats that are not always communicated clearly. LIME explanations can be unstable; small changes to the input produce noticeably different explanations. More troublingly, both LIME and SHAP can be defeated by adversarial models designed to appear fair when being audited while behaving differently in deployment. This is a serious problem for any regulatory framework that relies on these tools.

In neural networks, attention mechanisms assign weights to different parts of the input in a text model, which words mattered; in an image model, which regions. These weights are often presented as explanations for the model's behaviour. The intuition is appealing: if the model highlighted the patient's age and blood pressure when predicting cardiac risk, that seems like a meaningful explanation.

The problem is that this intuition may not be justified. Attention weights do not reliably correspond to causal influence on the output. In many cases, you can shuffle the attention weights substantially without changing the prediction.What these methods visualize may reflect the structure of the input data more than the reasoning of the model.

Even setting aside the technical limitations of specific methods, there is a more fundamental issue: transparency does not automatically produce accountability. An explanation can be technically valid but practically useless. It can be accurate on average but misleading in individual cases. And it can give the appearance of accountability without any of the substance.

Part of the problem is cognitive. Miller (2019) reviews a substantial body of work in cognitive science showing that humans prefer explanations that are causal, contrastive, and simple. We want to know not just why something happened, but why it happened rather than something else, and we want the answer in a form we can reason about. High-dimensional statistical outputs rarely satisfy these criteria naturally, and retrofitting them to do so often involves some distortion.

This is not an argument against transparency,it is an argument for understanding what transparency can and cannot do. Technical explainability is a necessary condition for accountability, but it is far from sufficient. The institutional and legal structures through which explanations become actionable matter just as much.

II. CONCLUSION

The transparency and explainability agenda in AI has made genuine progress. We have better tools than we did a decade ago, a more sophisticated understanding of what explanation means in different contexts, and a regulatory environment that is beginning to hold organizations accountable. That is worth acknowledging.

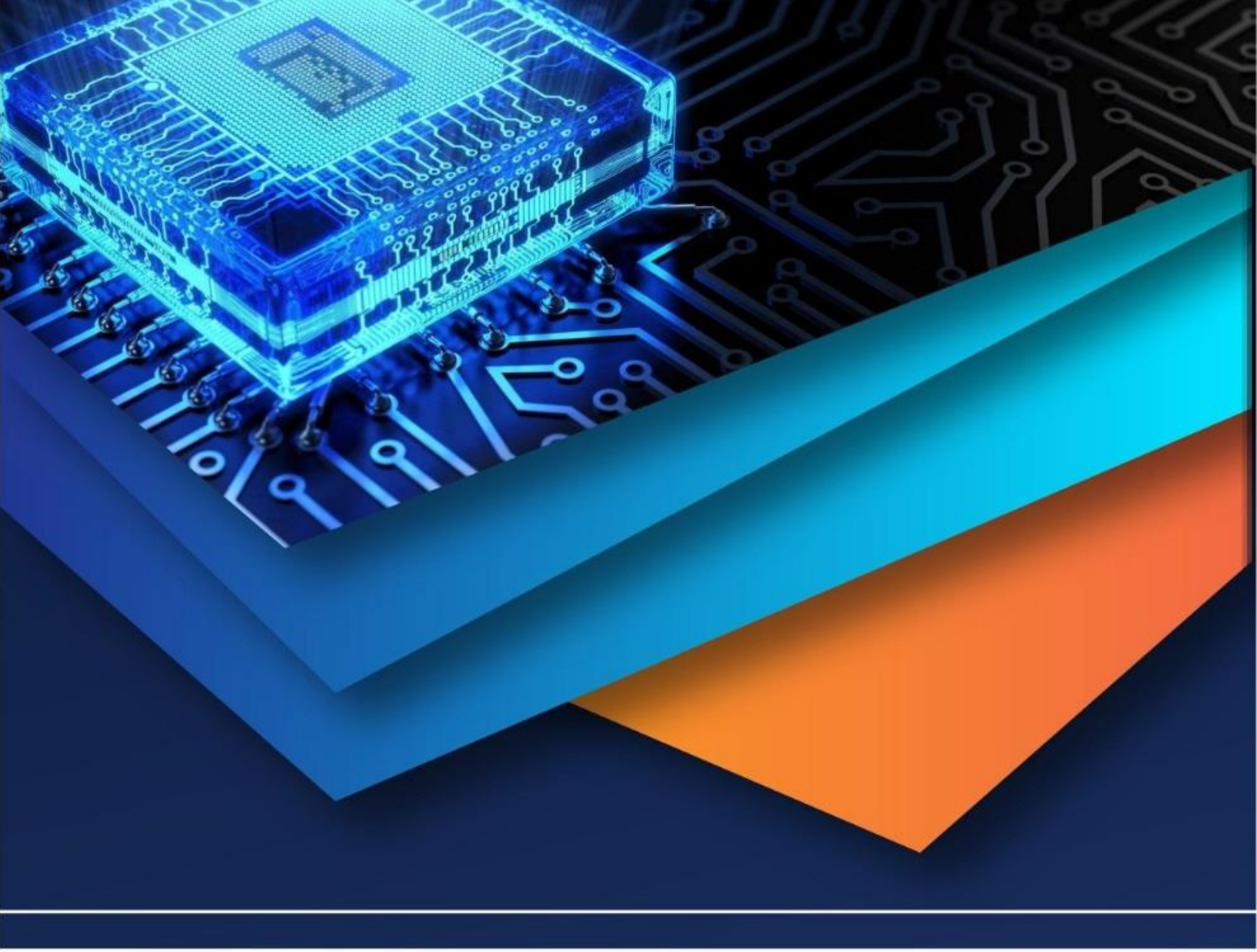
But the field has also developed some habits of thought that are worth questioning. There is a tendency to treat explainability as a property of models rather than a property of the relationship between models, people, and institutions. There is a tendency to assume that because an explanation can be generated, it is meaningful. And there is a tendency to invest heavily in technical solutions to problems that are, at their core, about power and accountability.

The people most affected by consequential AI systems those denied loans, flagged by predictive tools, screened out by automated hiring often have the least visibility into how those systems work and the fewest resources to contest their decisions. Addressing that asymmetry requires more than better visualization tools. It requires asking who gets to ask questions, who has the standing to demand answers, and what structures exist to ensure that explanations lead to consequences.

That is ultimately what responsible AI means: not just systems that can explain themselves, but systems embedded in institutions that are accountable for what those systems do.

REFERENCES

[1] Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems, 30.
[2] Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. Artificial Intelligence, 267, 1–38.
[3] Ribeiro, M. T., Singh, S., &Guestrin, C. (2016). "Why should I trust you?": Explaining the predictions of any classifier. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 1135–1144.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY