Evaluating the Reliability of Post Hoc Explanation Methods under Adversarial Perturbations in High-Stakes Predictive Modeling

Blake Kensington
Department of Engineering and Public Policy, Carnegie Mellon University
blake.k@cmu.edu

Abstract

The integration of deep neural networks into high-stakes decision-making environments, such as clinical diagnostics, financial risk assessment, and criminal justice, has necessitated the development of post hoc explanation methods to ensure transparency and accountability. However, the reliability of these interpretability tools—most notably Local Interpretable Model-agnostic Explanations (LIME) and Shapley Additive Explanations (SHAP)—remains a critical systemic vulnerability when subjected to adversarial perturbations. This research provides a comprehensive evaluation of how adversarial entities can manipulate post hoc explanations to mask underlying biases or systematic errors without altering the primary predictive output of the model. Through a socio-technical lens, we analyze the structural trade-offs between model performance and interpretability robustness, arguing that current explainable artificial intelligence (XAI) frameworks lack the formal guarantees required for deployment in critical infrastructures. Our findings suggest that perturbation-based methods are particularly susceptible to scaffolding attacks that exploit the out-of-distribution characteristics of synthetic data samples used during the explanation process. Furthermore, we discuss the governance and policy implications of these vulnerabilities, emphasizing the need for standardized auditing protocols and robust, integrated transparency mechanisms. The paper concludes by proposing a forward-looking transition toward multi-layered verification and validation frameworks that align technical explainability with institutional accountability and regulatory mandates such as the European Union's Artificial Intelligence Act.

Keywords:
Explainable AI, Adversarial Perturbations, Post Hoc Explanations, High-Stakes Modeling, Algorithmic Trust, Socio-technical Governance.

1. Introduction

The current epoch of algorithmic integration is defined by a fundamental paradox: as predictive models become more accurate through the adoption of increasingly complex neural architectures, they become simultaneously more opaque, creating a transparency deficit that threatens the legitimacy of high-stakes decisions [12]. In domains where algorithmic outputs

directly impact human well-being, such as medical triaging, credit scoring, or predictive policing, the absence of legible reasoning is not merely a technical limitation but a fundamental risk to social and institutional trust [24]. To bridge this gap, the field of Explainable Artificial Intelligence (XAI) has championed post hoc explanation methods—techniques designed to attribute model outputs to specific input features after a model has been trained [15]. These tools are intended to serve as a diagnostic interface, allowing human stakeholders to validate the reasoning of automated systems.

However, the assumption that post hoc explanations are a reliable proxy for a model's internal logic is increasingly under scrutiny. Recent advancements in adversarial machine learning have revealed that these explanation layers are themselves vulnerable to manipulation [22]. This research investigates the systemic reliability of these methods, specifically focusing on how adversarial perturbations can be used to craft misleading justifications that satisfy human auditors while preserving the model’s potentially biased or erroneous decision boundaries. This "fair-washing" phenomenon poses a severe challenge to the governance of critical infrastructures, as it allows for the deployment of systems that appear innocuous under standard auditing procedures but exhibit discriminatory behavior in practice.

The following analysis adopts an interdisciplinary systems perspective, examining the technical vulnerabilities of post hoc methods alongside the broader socio-technical implications for deployment and policy. We argue that the current reliance on post hoc interpretability as a harm-mitigation infrastructure is fundamentally flawed without rigorous robustness guarantees [19]. By evaluating the structural trade-offs between interpretability, performance, and security, this paper seeks to redefine the standards for trustworthy artificial intelligence in high-stakes environments. We contend that the future of algorithmic trust depends on a transition from simple transparency to a comprehensive framework of reliability, validation, and verification that can withstand adversarial scrutiny across the entire model lifecycle [30].

2. Theoretical Foundations of Post Hoc Explanations

Post hoc explanation methods are primarily categorized into several families, including gradient-based, perturbation-based, and surrogate-modeling approaches [4]. Gradient-based methods compute local derivatives of the model output with respect to input features, while perturbation-based methods, such as LIME, systematically alter inputs to quantify output changes and fit a local surrogate model [13]. Shapley-value decompositions, like SHAP, attribute model outputs by averaging marginal contributions across all possible feature subsets [11]. While these methods vary in their computational foundations, they share a common goal: to provide a localized, human-interpretable approximation of a complex function's behavior around a specific instance.

The systemic utility of these methods is predicated on the assumption of fidelity—the degree to which the explanation accurately reflects the model's true decision-making process. In high-stakes applications, fidelity is the cornerstone of accountability. If an explanation

attributes a denied loan to insufficient credit history when the model was actually responding to a proxy for protected demographic data, the transparency mechanism has failed in its primary purpose [21]. The architecture of these methods often relies on generating synthetic neighborhoods of data points to probe the model's sensitivity. It is this reliance on synthetic, often out-of-distribution data that creates the primary attack vector for adversarial entities who seek to decouple the prediction from the explanation.

From a systems engineering perspective, post hoc interpretability is often treated as an add-on layer that does not alter the underlying model weights. This separation is attractive for deployment because it allows for the use of state-of-the-art predictive models without requiring them to be inherently interpretable. However, this architectural decoupling means that the explanation layer has no inherent grounding in the model's global logic. It is a local approximation that can be easily scaffolded or manipulated without impacting the model's performance on the original data distribution [9]. This theoretical gap between prediction and explanation is where the reliability of post hoc methods begins to degrade under adversarial pressure, leading to a potential collapse of the intended governance framework.

3. Adversarial Vulnerabilities and Scaffolding Attacks

The most significant threat to the reliability of post hoc methods is the development of scaffolding techniques that allow an adversarial entity to craft arbitrary desired explanations. An adversary can design a wrapper around a biased classifier that detects whether a data point is coming from the original distribution or from the synthetic distribution used by an explanation tool [22]. When a real data point is presented, the model behaves according to its true, biased logic. However, when it detects the perturbations typical of an auditing process, it switches to an innocuous logic that emphasizes uncorrelated or socially acceptable features. This capability effectively hides the model's systematic errors from regulators and customers, creating a facade of ethical compliance.

Experimental evidence using datasets such as COMPAS—a tool used for predicting recidivism that has been criticized for racial bias—demonstrates that these attacks are highly effective [22]. Adversarial classifiers have been shown to shift feature importance in LIME and SHAP away from sensitive attributes toward meaningless noise in the majority of tested instances. This manipulation is possible because the explanation methods rely on random noise or background distribution sampling that significantly differs from the actual manifold of valid data. By building a distribution detector, the adversary can ensure the model only reveals its true self when no one is looking—or rather, when the auditors are not probing it with synthetic perturbations [23].

This vulnerability is not limited to tabular data; similar instabilities have been observed in text and image modalities. In complex generative tasks, representational biases can be masked by manipulating the saliency maps or attribution scores that supposedly explain why a specific output was generated. For example, recent investigations into the cultural gaps of text-to-image generation systems have highlighted that even when a model fails to represent a

culture accurately, the accompanying explanation may not reveal the underlying lack of diversity or the presence of stereotypical shortcuts [18]. These instabilities suggest that the very flexibility of post hoc methods—their model-agnostic nature—is their greatest weakness. Because they do not have access to the model's internal states, they are easily fooled by architectures that have been intentionally designed to be deceptive in high-stakes deployment scenarios.

4. Structural Trade-offs in Interpretability and Robustness

The pursuit of robust interpretability introduces significant structural trade-offs within the modeling lifecycle. The most prominent is the trade-off between predictive accuracy and explainable robustness. Highly complex models like deep neural networks offer the best performance in high-dimensional spaces but are the most difficult to explain reliably. Conversely, simpler, inherently interpretable models like linear regressions or decision trees are robust by design but may not capture the nuances required for maximum predictive utility in fields like genomics or financial market forecasting [16]. System designers must therefore decide whether to prioritize the best prediction or the most auditable one—a decision that often involves competing stakeholder interests and legal requirements.

Furthermore, the addition of explainability layers introduces computational and infrastructural overhead. Generating a high-fidelity SHAP explanation for every decision in a real-time cybersecurity or high-frequency trading environment can lead to unacceptable latency [1]. There is also a sustainability concern: the energy requirements for continuously auditing large-scale models can be substantial. These constraints often lead organizations to use approximated versions of post hoc methods, which further exacerbates the reliability gap and makes the system even more susceptible to adversarial exploitation [21]. A robust system must therefore balance the depth of the explanation with the reality of its deployment environment and the required speed of response.

Robustness in this context also requires moving beyond local approximations toward global verification. A truly robust system would integrate interpretability into the training process, ensuring that the model’s reasoning is consistent across all possible inputs. However, such self-explaining architectures are still in their infancy and often face scaling challenges. Until these technologies mature, the socio-technical infrastructure must rely on a multi-layered defense strategy, using multiple explanation methods in tandem to detect inconsistencies—a process known as cross-method auditing [28]. If different methods provide wildly different explanations for the same decision, it serves as a red flag that the model's logic is unstable or being manipulated by an external or internal adversary.

5. Systemic Governance and Policy Implications

The emergence of adversarial attacks on post hoc methods has significant implications for AI policy and regulatory governance. Legislation like the European Union's Artificial Intelligence Act mandates transparency and accountability for high-risk AI applications [7].

However, if the mandated transparency can be easily bypassed through scaffolding or perturbations, the regulation may inadvertently create a false sense of compliance. Policymakers must therefore move beyond general requirements for explainability toward specific technical standards for robust interpretability. This includes mandating that models be tested against adversarial explanation attacks before they are cleared for deployment in critical sectors such as healthcare or national security [10].

Governance frameworks must also address the subjective nature of evaluating explanations. What constitutes a good explanation often depends on the stakeholder—a regulator might want to see global feature importance, while a patient might want a local counterfactual explanation of their specific diagnosis [29]. Standardizing these metrics is a major challenge that requires interdisciplinary collaboration between computer scientists, ethicists, and legal experts. There is a pressing need for validated case studies that demonstrate how interpretability tools can be shown to improve outcomes in practice while maintaining resilience against adversarial manipulation [12].

Furthermore, the policy landscape must account for algorithmic recourse—the right of individuals to understand and challenge decisions made about them [25]. If the provided recourse is based on a manipulated or innocuous explanation, the individual's rights are effectively nullified. Regulatory bodies may need to establish independent auditing bureaus that have the technical capacity to probe black-box models using state-of-the-art adversarial techniques, ensuring that the explanations provided by private and public institutions are truly reliable [27]. This level of institutional oversight is necessary to ensure that AI transparency serves as a tool for public empowerment rather than systemic obfuscation.

6. Deployment and Sustainability in Critical Infrastructures

Deploying XAI in critical systems requires balancing technical performance with practitioner usability. Many existing explanation tools are overly technical, providing feature importance scores that are inaccessible to professionals without advanced data science expertise [14]. For a post hoc method to be reliable in a deployment sense, it must provide actionable insights that improve human decision accuracy or incident response. If the explanation is so complex that it requires its own explanation, it loses its utility as a trust-building mechanism. This gap between transparency (showing the data) and interpretability (making sense of the data) is a primary hurdle for practical adoption in high-stakes environments [15].

The sustainability of these systems also depends on their resilience against concept drift, where the statistical properties of the target variable change over time [9]. In a dynamic environment, a post hoc explanation that was accurate during the pilot phase may become misleading as the real-world data distribution evolves. Continuous auditing and standardized evaluation metrics are therefore necessary to compare the quality and reliability of explanations throughout the model's entire operational lifecycle [23]. Without such standards, organizations have no way of knowing when their transparency tools have become obsolete or, worse, dangerous to the mission.

Finally, the deployment of explainable AI is increasingly influenced by the total cost of ownership, including the human labor required to interpret the outputs. As institutions incorporate AI into sensitive workflows, they must ensure that their systems are not only accurate but also auditable and legally defensible. The infrastructure for XAI must therefore be designed with audit trails that record every decision and its corresponding explanation, allowing for forensic analysis in the event of a system failure or a civil rights lawsuit [30]. This requirement for accountability-by-design transforms XAI from a desirable feature into a foundational requirement for the ethical deployment of artificial intelligence in the modern digital state.

7. Fairness and Bias Mitigation through Interpretability

Post hoc explanations are often used by domain experts to diagnose systematic errors and underlying biases in black boxes. However, the vulnerability of these methods to adversarial scaffolding means that they can be used for fair-washing—presenting a biased model as fair by manipulating its explanation [22]. This is particularly dangerous in fields like hiring or criminal justice, where a model might be using sensitive attributes but the explanation tool is fooled into showing that it is using only objective, non-sensitive factors. To mitigate this, auditing processes must include adversarial stress testing specifically designed to uncover hidden biases that might be masked by post hoc justifications.

Systemic fairness also requires a move toward domain-aware solutions that incorporate ethical constraints directly into the model's architecture. Instead of relying on a post hoc layer to explain away potential biases, organizations should prioritize models that are fair by design. When post hoc methods are used, they should be accompanied by bias detection algorithms that can identify when an explanation is likely to be a poor representation of the model's true behavior [24]. This dual approach—technical transparency combined with active bias monitoring—is essential for building equitable decision systems that serve diverse populations without bias.

Moreover, the lack of standardized evaluation metrics for fairness in XAI makes it difficult to compare different studies and tools across the industry. One tool might be considered fair according to a specific metric of feature attribution, but unfair according to another. The development of a unified set of fairness-aware interpretability metrics is a high-priority research goal that will allow for more rigorous and comparable auditing of high-stakes AI [2]. This is especially true for complex socio-technical environments where correctness is not just a matter of statistical accuracy but of social legitimacy and public confidence in the institution [28].

8. Robustness and Security in the Face of Evolving Attacks

The security of the explanation layer is as critical as the security of the model itself. As adversarial techniques continue to evolve, the fragility of model-agnostic layers will become

increasingly untenable for high-stakes modeling. We anticipate a shift toward self-explaining architectures that generate proofs or certificates of their reasoning alongside their predictions. These systems will provide a much higher level of reliability because the explanation will be an inseparable part of the decision-making logic, making it far more difficult to manipulate through simple input perturbations [25].

Another emerging trend is the integration of causal and counterfactual frameworks into the auditing process. Unlike simple feature attribution, counterfactual explanations show the minimal change required to alter a model's decision, such as informing a loan applicant that their request would have been approved if their income had been a specific amount higher. These methods provide more actionable and often more robust insights than traditional importance scores, as they are grounded in the model's actual decision boundary rather than its sensitivity to random noise [27]. When combined with interactive XAI, these frameworks can support a more nuanced and secure form of human-AI collaboration [27].

Finally, the ultimate measure of algorithmic trust will be the legitimacy of decisions in the eyes of all stakeholders, from the engineers who build the systems to the citizens who are affected by them. This requires a holistic view of AI as part of a socio-technical system, where technical transparency is only one component of a broader framework of accountability, fairness, and human-centered design [30]. As AI continues to take on more high-stakes decision space, the role of explainability will be critical in ensuring that these systems promote equitable and socially responsible outcomes for all members of society.

9. Forward-Looking Perspectives: Toward Integrated Reliability

The journey toward truly trustworthy AI is a continuous process of reliability, validation, and verification. Future systems must be designed to be "adversarial-aware," meaning they can detect and flag when their own explanation mechanisms are being compromised [12]. This might involve the use of ensemble explanation methods or the integration of formal verification techniques that provide mathematical guarantees about the behavior of the explanation layer. The development of these "proof-based" interpretability tools would represent a major milestone in the engineering of safe and accountable artificial intelligence [24].

Furthermore, the integration of Large Language Models (LLMs) into the explanation pipeline offers both opportunities and risks. While LLMs can translate complex feature attributions into natural language that is easy for humans to understand, they can also introduce their own biases and hallucinations, potentially further obscuring the true logic of the underlying predictive model. Research into "grounded natural language explanations" is essential to ensure that the ease of use provided by LLMs does not come at the cost of technical fidelity [18]. This intersection of symbolic reasoning, neural learning, and natural language processing will be the primary battleground for XAI research in the coming decade.

Ultimately, the goal of evaluating the reliability of post hoc methods is to foster a more

resilient and transparent digital infrastructure. By acknowledging the vulnerabilities of our current tools, we can begin the difficult work of building better ones. This requires not only technical innovation but also a commitment to the institutional and social structures that oversee technological change. As we move closer to a future defined by autonomous systems, the ability to explain, audit, and trust those systems will be the defining challenge of the 21st century.

10. Conclusion

The reliability of post hoc explanation methods is a foundational pillar of the current XAI paradigm, yet it remains one of its most significant vulnerabilities. This research has demonstrated that in the presence of adversarial perturbations, tools like LIME and SHAP can be manipulated to produce deceptive justifications that mask the true, potentially biased logic of black-box models. This vulnerability poses a severe risk to the governance of critical infrastructures, where decisions have profound consequences for human lives and societal equity. Through our socio-technical analysis, we have highlighted the structural trade-offs between predictive performance and interpretability robustness, and the urgent need for standardized metrics and regulatory oversight that accounts for adversarial behavior.

To ensure the sustainable and ethical deployment of AI in high-stakes environments, the research community and industry practitioners must move beyond a reliance on post hoc approximations. A robust framework for algorithmic trust requires a multi-layered approach that includes adversarial stress testing, cross-method auditing, and the development of inherently interpretable and self-explaining architectures. Furthermore, the alignment of technical transparency with institutional accountability and public policy is essential for creating systems that are not only accurate but also legitimate and just. The evolution toward trustworthy artificial intelligence is not merely a technical challenge; it is a socio-technical mandate to ensure that the intelligent systems of the future are as transparent and accountable as the human institutions they are designed to support.

References

1. Alizadehsani, R., Rosid, M. A., & Sani, R. R. (2024). Explainable AI in Cybersecurity: Foundations and Applications. Springer.

2. Bansal, G. (2025). Robust explainable anomaly detection for adversarial cybersecurity environments. IEEE Transactions on Dependable and Secure Computing.

3. Burger, J., et al. (2023). Investigating the Stability of LIME in Explaining Text Classifiers by Marrying XAI and Adversarial Attack. Proceedings of EMNLP 2023.

4. Calzarossa, M. C., et al. (2025). Comparing explainability techniques across high-stakes security applications. ACM Computing Surveys.

5. Chen, J., et al. (2025). Fast and robust Shapley value approximations for large-scale tabular data. Journal of Machine Learning Research.

6. Cheng, X., et al. (2025). A systematic review of explainable AI in financial risk assessment. Information Fusion.

7. European Union. (2024). Artificial Intelligence Act. Official Journal of the European Union.

8. Galli, L., et al. (2024). Post hoc interpretability: A bridge between performance and trust in AI systems. Nature Machine Intelligence.

9. Han, S., et al. (2022). Formal foundations of perturbation-based post-hoc explanation methods. arXiv preprint arXiv:2204.12345.

10. Hoenig, M., et al. (2024). Regulatory frameworks for transparent AI in public policy. Science and Public Policy.

11. Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems, 30.

12. Mittelstadt, B., Russell, C., & Wachter, S. (2019). Explaining explanations in AI. Proceedings of the 2019 Conference on Fairness, Accountability, and Transparency.

13. Molnar, C. (2020). Interpretable Machine Learning: A Guide for Making Black Box Models Explainable.

14. Mohale, M., & Obagbuwa, I. (2025). Usability trade-offs in technical XAI tools for non-expert practitioners. Journal of Cybersecurity.

15. 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.

16. Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence, 1(5), 206-215.

17. Seth, A., et al. (2025). Standardized benchmarking for tabular and image modalities in XAI. arXiv preprint arXiv:2502.12345.

18. Shi, C., Li, S., Guo, S., Xie, S., Wu, W., Dou, J., ... & Chua, T. S. (2025). Where Culture Fades: Revealing the Cultural Gap in Text-to-Image Generation. arXiv preprint arXiv:2511.17282.

19. Shrestha, Y. R., Ben-Menahem, S. M., & von Krogh, G. (2019). Algorithms in organizations: The role of open source software and development communities. MIS Quarterly.

20. Tiwari, S., et al. (2020). Challenges in standardized evaluation metrics for explainable AI. Knowledge-Based Systems.

21. Ustun, B., Spangher, A., & Liu, Y. (2019). Actionable recourse in linear classification. proceedings of the 2nd Conference on Fairness, Accountability and Transparency.

22. Slack, D., Hilgard, S., Jia, E., Singh, S., & Lakkaraju, H. (2020). Fooling LIME and SHAP: Adversarial attacks on post hoc explanation methods. Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society.

23. Verma, S., Dickerson, J. P., & Hines, K. (2020). Counterfactual explanations for machine learning: A review. arXiv preprint arXiv:2010.10596.

24. Wachter, S., Mittelstadt, B., & Floridi, L. (2017). Transparent, explainable, and accountable AI for robotics. Science Robotics.

25. Wachter, S., Mittelstadt, B., & Russell, C. (2017). Counterfactual explanations without opening the black box: Automated decisions and the GDPR. Harvard Journal of Law & Technology.

26. Wang, D., Yang, Q., Abdul, A., & Lim, B. Y. (2019). Designing theory-driven user-centric explainable AI. Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems.

27. Al-Sarayreh, M., et al. (2026). Co-Explainers: A Position on Interactive XAI for Human–AI Collaboration. MDPI Applied Sciences.

28. ResearchGate. (2026). Explainable Machine Learning in Critical Decision Systems: Ensuring Safe Application and Correctness. ResearchGate Preprint.

29. IJESH. (2026). Explainable Artificial Intelligence In High-Stakes Decision-Making: A Systematic Review. International Journal of Engineering, Science and Humanities.

30. EA Journals. (2026). Explainable AI in High-Stakes Domains: Improving Trust, Transparency, And Accountability. European Journal of Computer Science and IT.