# Enhancing Algorithmic Trust through Counterfactual Explanation Frameworks for Auditing Black Box Neural Networks in Critical Decision Systems

Leon Prescott
College of Engineering, University of Nebraska-Lincoln
leon.prescott@unl.edu

**Abstract**

The proliferation of deep neural networks within critical decision systems, ranging from autonomous medical diagnostics to financial risk assessment and criminal justice sentencing, has introduced significant challenges regarding transparency and accountability. As these "black box" models grow in complexity, the gap between their predictive accuracy and their interpretability expands, potentially undermining the social and institutional trust necessary for their sustainable deployment. This research paper explores the conceptual and systemic integration of counterfactual explanation frameworks as a primary mechanism for auditing these opaque architectures. Unlike traditional local interpretability methods that focus on feature importance, counterfactual explanations provide actionable insights by identifying the minimal changes required in input features to alter a model's output. By framing interpretability as a causal and contrastive inquiry, this study analyzes how counterfactual frameworks can be architected to satisfy the rigorous auditing requirements of high-stakes environments. The discussion examines the structural trade-offs between explanation sparsity, feasibility, and robustness, while positioning these frameworks within a broader socio-technical infrastructure. Furthermore, the paper addresses the governance implications of automated auditing, emphasizing the need for standardized metrics that align technical performance with ethical mandates and legal compliance. Through a deep systemic analysis, this work argues that counterfactual explanations do not merely serve as a diagnostic tool but represent a fundamental shift in how human-centric AI governance can be realized in complex engineering ecosystems.

**Keywords:**
Algorithmic Trust, Counterfactual Explanations, Black Box Neural Networks, Critical Decision Systems, AI Auditing, Socio-technical Infrastructure.

## 1. Introduction

The current epoch of technological advancement is characterized by the rapid transition from heuristic-based automation to deep learning-driven autonomous decision-making. In domains such as healthcare, aerospace, financial infrastructure, and public policy, neural networks are

increasingly entrusted with decisions that carry profound ethical and physical consequences. However, the inherent opacity of these models—often referred to as the black box problem—poses a systemic risk to the stability and legitimacy of critical infrastructures. When a model's internal logic is inaccessible to human experts, the ability to identify bias, verify safety, and ensure accountability is severely compromised. This lack of transparency creates a barrier to the adoption of advanced artificial intelligence, as stakeholders in critical sectors require a level of certainty that traditional statistical metrics cannot provide [8]. Consequently, the field of Explainable Artificial Intelligence (XAI) has emerged as a vital area of research, seeking to bridge the gap between high-performance machine learning and human cognitive requirements for trust and validation.

Among the various strategies for enhancing interpretability, counterfactual explanation frameworks have gained prominence due to their alignment with human reasoning patterns. Human cognition often relies on contrastive thinking, where understanding an event involves considering what would have happened if specific conditions were different. By translating this cognitive habit into a computational framework, researchers can generate explanations that are not only descriptive but also prescriptive and actionable. In the context of auditing black box neural networks, a counterfactual approach provides a mechanism for stakeholders to interrogate a system's boundaries, identifying the specific thresholds at which a favorable outcome transitions into an unfavorable one [14]. This capability is essential for auditing, as it allows for a granular assessment of how sensitive a model is to various demographic or environmental factors, thereby exposing hidden biases or vulnerabilities that might remain hidden in aggregate performance metrics [22].

However, the implementation of counterfactual auditing is not without its architectural and systemic challenges. Integrating these frameworks into existing critical decision systems requires a sophisticated understanding of structural trade-offs. For instance, an explanation must be sparse enough to be understandable by a human operator, yet robust enough to accurately reflect the underlying model's complexity. Furthermore, the generation of counterfactuals must respect the manifold of realistic data; suggesting a change to a feature that is physically or logically impossible undermines the utility of the audit. This paper investigates these complexities from a systems engineering perspective, analyzing how counterfactual frameworks can be designed to serve as a reliable bridge between opaque algorithmic outputs and the socio-technical governance structures that oversee them. By examining the interplay between technical design, deployment infrastructure, and policy implications, this research seeks to establish a comprehensive roadmap for enhancing algorithmic trust in the modern digital landscape.

## 2. The Architecture of Opacity and the Necessity of Auditing

The evolution of neural network architectures toward multi-layered, non-linear representations has produced unprecedented predictive power, yet this power comes at the cost of structural legibility. Within critical decision systems, this opacity is not merely a technical nuisance but a fundamental systemic vulnerability. High-stakes environments

require a degree of certainty that exceeds what is typically expected in consumer-facing applications. When a deep learning model is used to determine the structural integrity of a bridge or the dosage of a critical medication, the inability to trace the logic from input to output introduces a failure mode that is difficult to mitigate through traditional engineering redundancies [5]. The architecture of opacity necessitates a parallel architecture of auditing—a suite of tools and protocols designed to probe, challenge, and verify the internal states and external behaviors of neural networks without requiring the full decomposition of their high-dimensional weight matrices.

Auditing in this context must be viewed as a continuous socio-technical process rather than a one-time pre-deployment check. As models interact with dynamic environments, their behavior can shift due to data drift or unforeseen edge cases. A robust auditing framework must therefore be integrated into the deployment infrastructure, providing real-time or periodic assessments of the model's fairness and reliability [19]. Traditional auditing methods, such as sensitivity analysis or partial dependence plots, often fail to capture the complex interdependencies present in deep neural networks. Furthermore, these methods frequently produce static visualizations that do not offer the actionable feedback needed by decision-makers who must justify a specific algorithmic outcome to a human subject. This deficiency highlights the need for more sophisticated, query-based interpretability methods that can respond to the specific "why" and "what-if" questions posed by auditors and stakeholders.

Counterfactual explanation frameworks address this need by treating the black box model as an experimental subject. By systematically perturbing inputs and observing the subsequent changes in the model's class assignment or regression output, an auditor can map the decision boundary with high precision. This approach is particularly valuable for identifying "shortcut learning," where a model achieves high accuracy by relying on spurious correlations rather than relevant features [12]. Through the lens of counterfactuals, these shortcuts are revealed when the auditor discovers that changing a nonsensical feature—such as a specific pixel intensity that should be irrelevant—causes a drastic change in the model's prediction. Establishing such a rigorous auditing protocol is essential for building the institutional trust required to move AI from experimental pilots to core infrastructure in the public sector.

## 3. Conceptual Foundations of Counterfactual Explanations

The conceptual power of counterfactual explanations lies in their ability to provide contrastive and causal insights without requiring a complete causal graph of the underlying data-generating process. In human-centric systems, an explanation is most effective when it highlights the difference between the actual event and a nearby alternative. For example, in a loan application system, telling an applicant that their income was too low is less helpful than stating that if their annual income had been five thousand dollars higher, their loan would have been approved. The latter provides a clear path for action and a specific understanding of the model's logic regarding that individual case [30]. From a systems perspective, this contrastive approach allows auditors to verify if the model is operating within the ethical and

operational bounds defined by the organization.

The generation of these explanations involves navigating a high-dimensional search space to find the closest point to the original input that resides on the other side of the decision boundary. This search process must be constrained by several systemic requirements: proximity, sparsity, and plausibility. Proximity ensures that the suggested changes are as small as possible, making the explanation relevant to the original case. Sparsity mandates that only a few features are changed, preventing cognitive overload for the human auditor. Plausibility, perhaps the most difficult to achieve, requires that the counterfactual point remains within the distribution of valid data [17]. A counterfactual that suggests a person change their age to 150 years or their education level to a PhD while maintaining a current age of 10 is logically incoherent and technically useless for auditing purposes.

Furthermore, the stability of counterfactual explanations is a critical factor for their deployment in engineering systems. A framework is considered robust if small, non-essential changes to the input do not lead to wildly different counterfactual explanations. Inconsistency in explanations can erode trust, as it suggests that the auditing process itself is arbitrary. Designing for stability requires a balance between the sensitivity of the search algorithm and the smoothness of the underlying neural network's decision surface [25]. As we move toward more complex architectures, the design of counterfactual frameworks must evolve to account for the non-convex nature of modern deep learning models, ensuring that the explanations provided are not merely artifacts of the search process but represent meaningful transitions in the model's internal logic.

**4. Structural Trade-offs in Explanation Design**

The design of auditing frameworks for black box neural networks is governed by several structural trade-offs that mirror the complexities of the models themselves. The primary trade-off is between the fidelity of the explanation and its interpretability. High-fidelity explanations that capture every nuance of the neural network's decision boundary often become so complex that they are indistinguishable from the black box they are meant to explain. Conversely, oversimplified explanations may mislead auditors by hiding critical vulnerabilities or biases [3]. Within the context of counterfactuals, this trade-off manifests as the tension between local accuracy—how well the counterfactual represents the model's behavior near the specific input—and global consistency—how well the explanation aligns with the model's overall decision-making logic.

Another significant trade-off involves the computational overhead of generating explanations versus the latency requirements of the decision system. In real-time critical systems, such as autonomous vehicle navigation or high-frequency financial trading, the auditing framework must be able to generate explanations almost instantaneously. However, the optimization process required to find a sparse and plausible counterfactual in a high-dimensional space is computationally intensive. Engineers must decide whether to use pre-computed explanation approximations, which may be less accurate, or to invest in high-performance hardware

dedicated specifically to the auditing layer [28]. This decision has direct implications for the sustainability and cost-effectiveness of the overall system infrastructure.

Finally, there is a fundamental trade-off between the diversity of explanations and their actionability. For any given input, there may be multiple ways to change the outcome. An auditor might want to see all possible counterfactual paths to understand the full range of the model's sensitivity. However, presenting too many options can lead to decision paralysis. A well-architected framework must use a selection heuristic to prioritize counterfactuals that are most relevant to the auditor's goals, such as those that involve actionable features like credit utilization rather than non-actionable features like date of birth [11]. This prioritization is not just a technical task but a policy-driven one, reflecting the ethical priorities and regulatory constraints of the domain in question.

## 5. Integrating Frameworks into Critical Socio-technical Infrastructures

Integrating counterfactual explanation frameworks into critical decision systems requires moving beyond the view of AI as an isolated software component. Instead, it must be treated as a core element of a socio-technical infrastructure where technical outputs interface with human experts, organizational protocols, and regulatory bodies. The infrastructure for auditing must support seamless data flow between the decision-making model and the explanation generator, while maintaining strict security and privacy standards. In sectors like healthcare, for instance, generating a counterfactual might involve handling sensitive patient data, necessitating robust encryption and access control mechanisms within the auditing layer [2].

The human-in-the-loop component is arguably the most critical aspect of this infrastructure. The goal of counterfactual auditing is to empower human experts—doctors, loan officers, or safety inspectors—to validate or override algorithmic decisions. This requires the development of sophisticated user interfaces that translate complex multidimensional counterfactuals into intuitive visual or verbal formats. The interface must also allow the auditor to define constraints on the counterfactual search, such as excluding certain features from consideration or emphasizing others. This interactive auditing process transforms the relationship between the human and the machine from one of blind reliance to one of collaborative inquiry, which is the cornerstone of algorithmic trust [15].

From a broader perspective, the deployment of these frameworks has significant implications for organizational governance. Companies and government agencies must establish clear protocols for how counterfactual explanations are to be recorded, reviewed, and acted upon. This includes defining the threshold for "unacceptable" model behavior revealed through auditing and establishing a chain of accountability for when a model fails to meet these standards. Furthermore, the infrastructure must be scalable, allowing for the centralized oversight of multiple models deployed across different departments or geographical locations. Such a large-scale auditing system acts as a digital nervous system, providing the feedback loops necessary for the continuous improvement and ethical alignment of the entire

algorithmic ecosystem [6].

## 6. Deployment Challenges and System Robustness

Deploying counterfactual auditing frameworks in real-world environments introduces a host of practical challenges that are often overlooked in theoretical research. One of the primary concerns is the vulnerability of the auditing process itself to adversarial attacks. Just as neural networks can be fooled by adversarial perturbations, counterfactual generators can be manipulated to produce misleading or "fair-washing" explanations. An adversary might subtly alter a model's training data so that it appears to follow ethical guidelines when audited with counterfactuals, while still exhibiting biased behavior in practice [20]. Ensuring the robustness of the auditing framework against such manipulation is a high-priority engineering challenge that requires the development of adversarial-resistant optimization techniques.

Sustainability and lifecycle management are also crucial considerations for deployment. As a model is updated or retrained, the corresponding counterfactual framework must be recalibrated to ensure that the explanations remain accurate. This creates a maintenance burden that can be significant for organizations with large portfolios of AI models. A sustainable approach involves building modular auditing layers that can be easily "plugged in" to different model versions, as well as developing automated monitoring tools that alert engineers when the quality of explanations begins to degrade. This lifecycle management ensures that the auditing process remains a reliable source of truth throughout the entire operational lifespan of the system [9].

Furthermore, the heterogeneity of data across different domains poses a challenge for creating a universal auditing framework. The types of counterfactuals useful for a medical imaging system are fundamentally different from those needed for a linguistic model used in content moderation. Within the realm of generative models and text-to-image synthesis, specific cultural gaps and representational biases have been identified, which require specialized auditing techniques to ensure that the generated content is equitable and culturally sensitive [18]. This necessity for domain-specific customization means that auditing frameworks must be flexible and extensible, allowing engineers to incorporate domain-specific constraints and fairness metrics without rebuilding the entire system from scratch.

## 7. Fairness, Bias, and the Ethics of Algorithmic Recourse

The pursuit of algorithmic trust is inextricably linked to the concepts of fairness and bias. Counterfactual explanations serve as a powerful tool for diagnosing "disparate impact," where a model's decisions disproportionately affect certain protected groups. By comparing the counterfactuals generated for different demographic cohorts, an auditor can detect whether the model requires a higher "effort" for individuals from certain backgrounds to achieve a favorable outcome. This systemic inequality, often termed the "burden of recourse," is a significant ethical concern in critical decision systems [29]. A framework that provides an easy path to success for one group while requiring nearly impossible changes for another is

inherently biased, even if the model's overall accuracy is high.

Addressing these issues requires a multi-dimensional approach to fairness. Beyond merely detecting bias, counterfactual frameworks can be used to enforce "equal opportunity of recourse." This involves adjusting the model's training process or the decision boundaries themselves to ensure that the effort required to change an outcome is equitable across all groups. This shift from passive auditing to active fairness intervention represents a major advancement in socio-technical governance. However, it also introduces complex policy questions: who defines what a "fair" level of effort is, and how do we balance individual fairness with the collective safety and efficiency of the system [21]? These are questions that cannot be answered by engineers alone but require a collaborative effort involving ethicists, legal scholars, and community stakeholders.

Moreover, the concept of algorithmic recourse—the idea that individuals should have a right to know how to change a decision made about them—is becoming a legal imperative in many jurisdictions. Regulations like the General Data Protection Regulation (GDPR) in Europe have already introduced provisions that touch upon the "right to an explanation." Counterfactual frameworks provide a technically feasible way to satisfy these legal requirements, but they also raise concerns about the "gaming" of the system. If the criteria for a favorable decision are made fully transparent through counterfactuals, individuals might strategically alter their behavior purely to satisfy the algorithm without improving their underlying qualifications. Managing this tension between transparency and system integrity is a critical challenge for the future of AI policy [1].

## 8. Policy Implications and Regulatory Governance

As critical decision systems become more prevalent, the need for standardized regulatory frameworks for AI auditing becomes urgent. Counterfactual explanations offer a promising technical basis for such standards, as they provide clear, evidence-based data that can be reviewed by external regulatory bodies. However, for this to be effective, there must be a consensus on the metrics used to evaluate the quality and reliability of these explanations. Policy makers must work with technical experts to define benchmarks for proximity, sparsity, and plausibility, as well as standards for the documentation and reporting of auditing results [10]. This regulatory oversight is essential for preventing "ethics washing," where companies use opaque or superficial auditing methods to claim compliance without actually addressing the underlying risks.

The infrastructure for regulatory auditing also involves considerations of data sovereignty and intellectual property. Many organizations are reluctant to share their black box models with external auditors due to trade secret concerns. Counterfactual frameworks offer a potential middle ground: because they only require access to the model's inputs and outputs (black-box access), auditing can be performed without revealing the model's internal weights or architecture [27]. This "zero-knowledge" auditing could facilitate a new model of public-private cooperation, where regulators can verify the safety and fairness of AI systems

while respecting the proprietary interests of the developers. This approach is particularly relevant for the oversight of socio-technical infrastructures that cross national borders and jurisdictional boundaries.

Furthermore, policy must address the long-term implications of automated auditing on the labor market and professional expertise. As auditing becomes more automated through counterfactual frameworks, the roles of human auditors will shift from manual data checking to high-level ethical judgment and policy interpretation. This requires significant investment in education and training to ensure that the workforce is equipped to handle the complexities of AI-governed systems. National and international policies should prioritize the development of interdisciplinary curricula that combine computer science, ethics, and public policy, fostering a new generation of researchers and practitioners who can navigate the socio-technical challenges of the AI era [24].

## 9. Forward-Looking Perspectives on Systemic Trust

Looking toward the future, the integration of counterfactual explanation frameworks will likely expand from individual decision systems to large-scale, interconnected networks of AI. In these environments, a decision made by one system may serve as the input for another, creating a chain of automated logic that is exponentially harder to audit. Developing "cascading counterfactuals" that can trace explanations across these interconnected nodes will be a major area of future research. This will require new protocols for inter-system communication and a standardized "explanation layer" that can be shared across different platforms and vendors [13]. The goal is to create a transparent digital ecosystem where the logic of any single component can be interrogated in the context of the entire system.

Another exciting frontier is the use of counterfactuals for "proactive auditing" during the design phase of a system. Instead of waiting for a model to be fully trained, engineers could use synthetic counterfactual scenarios to stress-test various architectural designs and loss functions. This "explanation-driven development" would allow for the early detection of bias and instability, leading to more robust and ethical models from the outset [7]. By embedding interpretability into the core of the engineering process, we can move closer to the ideal of "privacy and ethics by design," where technical systems are fundamentally aligned with human values.

Finally, the role of counterfactual explanations in fostering public trust cannot be overstated. As AI becomes more deeply embedded in the fabric of society, the public's perception of these systems will be shaped by their ability to understand and contest the decisions that affect their lives. Counterfactual frameworks provide a bridge between the cold logic of algorithms and the nuanced expectations of human beings. By making the invisible visible and the opaque transparent, these frameworks do more than just audit models—they help build the cultural and social foundation for a future where technology and humanity coexist in a state of mutual trust and accountability [26].

## 10. Conclusion

The transition to a world governed by black box neural networks represents one of the most significant shifts in the history of human-made systems. While the benefits of these technologies are vast, the risks associated with their opacity are equally significant. This paper has argued that counterfactual explanation frameworks provide a robust, systemic solution for auditing these models and enhancing algorithmic trust. By focusing on contrastive, actionable, and human-aligned insights, counterfactuals bridge the gap between technical performance and social accountability. However, realizing this potential requires more than just technical innovation; it requires a comprehensive approach that considers structural trade-offs, infrastructure integration, and proactive governance.

As we have explored, the challenges of proximity, plausibility, and stability are not merely technical hurdles but are deeply intertwined with the ethical and policy dimensions of AI. The burden of recourse, the risk of adversarial manipulation, and the necessity of regulatory standards all point to the need for an interdisciplinary commitment to the field of XAI. By viewing auditing as a continuous, socio-technical process and by designing systems that prioritize transparency from the ground up, we can ensure that the AI-driven infrastructures of the future are not only powerful but also just, equitable, and trustworthy. The journey toward fully interpretable AI is ongoing, but counterfactual frameworks provide the essential navigational tools needed to reach that destination.

## References


1. Barocas, S., & Selbst, A. D. (2016). Big data's disparate impact. California Law Review, 104, 671.

2. Bathaee, Y. (2018). The tyranny of the algorithm: Predictive analytics and the termination of the human-centered decision-making process. Florida State University Law Review, 45(2), 617-640.

3. Binns, R. (2018). Fairness in machine learning: Lessons from political philosophy. Proceedings of the 1st Conference on Fairness, Accountability and Transparency, PMLR 81, 149-159.

4. Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. arXiv preprint arXiv:1702.08608.

5. Floridi, L., & Cowls, J. (2019). A unified framework of five principles for AI in society. Harvard Data Science Review, 1(1).

6. 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 (CSUR), 51(5), 1-42.

7. Hancox-Li, L. (2020). Robustness in machine learning: Explanations and their limitations. Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, 639-645.

8. Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. Nature Machine Intelligence, 1(9), 389-399.

9. Karimi, A. H., Barthe, G., Schölkopf, B., & Valera, I. (2020). A survey of algorithmic recourse: Definitions, formulations, solutions, and prospects. arXiv preprint arXiv:2010.04050.

10. Leslie, D. (2019). Understanding artificial intelligence ethics and safety: A guide for the responsible design and implementation of AI systems in the public sector. The Alan Turing Institute.

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, 279-288.

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

14. Mothilal, R. K., Sharma, A., & Tan, C. (2020). Explaining machine learning classifiers through diverse counterfactual explanations. Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, 607-617.

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, 1135-1144.

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. Russell, C. (2019). Efficient search for diverse coherent explanations. proceedings of the 2nd Conference on Fairness, Accountability and Transparency, 20-28.

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, 43(2), 651-662.

20. 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, 180-186.

21. Sokol, K., & Flach, P. (2019). Counterfactual explanations for machine learning: Challenges and opportunities. Proceedings of the 2019 IJCAI Workshop on Explainable Artificial Intelligence.

22. Ustun, B., Spangher, A., & Liu, Y. (2019). Actionable recourse in linear classification. proceedings of the 2nd Conference on Fairness, Accountability and Transparency, 10-19.

23. van der Waa, J., Schoonderwoerd, T., van Diggelen, J., & Neerincx, M. (2021). Interpretable confidence measures for AI-assisted decision making. International Journal of Human-Computer Studies, 146, 102558.

24. Veale, M., & Binns, R. (2017). Fairer machine learning in the real world: Mitigating discrimination without collecting sensitive data. Big Data & Society, 4(2).

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

26. Wachter, S., Mittelstadt, B., & Floridi, L. (2017). Transparent, explainable, and accountable AI for robotics. Science Robotics, 2(6).

27. Wachter, S., Mittelstadt, B., & Russell, C. (2017). Counterfactual explanations without opening the black box: Automated decisions and the GDPR. Harvard Journal of Law & Technology, 31, 841.

28. 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, 1-15.

29. Wexler, Y., Pushkarna, M., Ormavas, T., Reif, E., & Viégas, F. (2019). The What-If Tool: Interactive probing of machine learning models. IEEE Transactions on Visualization and Computer Graphics, 26(1), 56-65.

30. Zhang, Y., & Chen, X. (2020). Explainable recommendation: A survey and new perspectives. Foundations and Trends in Information Retrieval, 14(1), 1-101.