International Journal of Engineering, Science and Humanities
An international peer reviewed, refereed, open-access journal
Impact Factor 8.3 www.ijesh.com ISSN: 2250-3552

Explainable Artificial Intelligence In High-Stakes Decision-Making: A Systematic Review Of Methods, Applications, And Challenges

C. Sumanth
Research Scholar, Department of Computer Science, North East Christian University
Dr. Kritesh Sharan
Associate Professor, Department of Computer Science, North East Christian University

Abstract

Explainable Artificial Intelligence (XAI) has become a vital area of research as we see more complex machine learning models being used in critical decision-making fields like healthcare, finance, criminal justice, and public policy. While these advanced AI systems often deliver impressive predictive accuracy, their black-box nature raises important questions about transparency, accountability, fairness, and trust—especially when their decisions can significantly affect people's lives and societal outcomes. This systematic review dives into the current state of explainable AI methods, their practical applications, and the hurdles we face when trying to implement XAI in crucial settings. The review brings together existing research on key explainability strategies, such as intrinsic interpretable models, post-hoc explanation techniques like LIME and SHAP, saliency-based visualization methods, and the newer counterfactual and causal explanation frameworks. Additionally, the paper showcases how XAI enhances decision support systems by boosting interpretability, aiding in bias detection, and building user trust in sensitive applications. It pays special attention to how these methods are applied in specific sectors, including clinical diagnosis, credit scoring, fraud detection, and risk assessment tools in the legal system. Despite the progress made, several challenges still need to be addressed, such as balancing model accuracy with interpretability, the subjective nature of evaluating explanations, scalability issues in real-time systems, and the potential for misleading or adversarial explanations. The review also touches on ethical, legal, and regulatory aspects, stressing the importance of standardized evaluation metrics and governance frameworks to ensure responsible AI deployment. This systematic review highlights just how crucial explainable AI is for building decision-making systems that are trustworthy, transparent, and ethically sound. It also points out future research paths that could lead to more robust, human-centered, and domain-aware solutions for explainability.

Keywords: Explainable Artificial Intelligence, High-Stakes Decision-Making, Interpretability, Transparency, Trustworthy AI, LIME, SHAP, Fairness, Responsible AI, Systematic Review

International Journal of Engineering, Science and Humanities
An international peer reviewed, refereed, open-access journal
Impact Factor 8.3 www.ijesh.com ISSN: 2250-3552

Introduction

Artificial Intelligence (AI) and machine learning (ML) technologies have quickly changed the way decisions are made in many important fields, like healthcare, finance, criminal justice, and autonomous systems. These technologies are increasingly trusted for tasks such as diagnosing diseases, scoring credit, detecting fraud, and predicting recidivism. While today's ML models—especially deep learning and ensemble methods—often boast impressive predictive accuracy, they tend to function as black boxes, leaving us in the dark about how decisions are reached. This lack of clarity is particularly concerning in high-stakes situations, where wrong or biased outcomes can have serious repercussions for individuals and society as a whole.

As the call for transparency and accountability grows louder, we see the rise of Explainable Artificial Intelligence (XAI), which seeks to make AI decisions more understandable for humans. This clarity is especially crucial in sensitive areas where trust, ethical responsibility, and legal compliance are paramount. As Lipton (2018) points out, interpretability isn't just a technical detail; it's essential for the responsible use of AI. Likewise, Doshi-Velez and Kim (2017) emphasize the importance of robust evaluation frameworks to ensure that interpretability methods truly enhance human understanding.

Explainable AI techniques can generally be divided into two categories: intrinsic interpretability approaches, where models are built to be transparent from the start, and post-hoc explanation methods, which try to make sense of complex models after they've been trained. Well-known post-hoc methods like LIME (Ribeiro et al., 2016) and SHAP (Lundberg & Lee, 2017) have gained a lot of attention for their ability to offer both local and global insights into how models behave.

Take the European Union's General Data Protection Regulation (GDPR), for example; it really highlights how crucial transparency is in automated decision-making systems (European Union, 2016). On a similar note, the National Institute of Standards and Technology has put forward risk management frameworks to help ensure that AI systems are deployed in a trustworthy manner (NIST, 2023).

This systematic review takes a closer look at the current methods, applications, and challenges surrounding explainable AI, especially in high-stakes decision-making scenarios. By bringing together insights from recent literature, the review seeks to pinpoint the strengths and weaknesses of existing XAI approaches while also shining a light on future research paths for creating robust, fair, and human-centered explainability solutions.

Explainable Artificial Intelligence (XAI) is all about creating methods and frameworks that help us understand, clarify, and interpret the results produced by AI systems. As machine learning models grow more intricate—especially with the advent of deep neural networks and

International Journal of Engineering, Science and Humanities
An international peer reviewed, refereed, open-access journal
Impact Factor 8.3 www.ijesh.com ISSN: 2250-3552

ensemble techniques—their decision-making processes can become quite murky. This lack of clarity raises significant issues in critical areas where decisions need to be justified, trusted, and held to ethical standards.

At the heart of XAI is the broader idea of model interpretability, which refers to how well a human can grasp the inner workings or reasoning behind a model's predictions. Lipton (2018) points out that interpretability isn't just one single trait; it's a collection of desirable features like transparency, simplicity, and explainability. In crucial applications, being able to interpret models is vital not just for troubleshooting but also for ensuring fairness and meeting regulatory requirements.

XAI methods typically fall into two main categories: intrinsically interpretable models and post-hoc explanation techniques. Intrinsically interpretable models—like decision trees, linear regression, and rule-based systems—are built to be understandable from the get-go. They offer clear insights into how different input features influence predictions, making them especially valuable in high-stakes decision-making scenarios (Rudin, 2019). However, it's worth noting that these models might sometimes compromise on predictive accuracy compared to more complex black-box algorithms.

2. Concept and Foundations of Explainable Artificial Intelligence (XAI)

In contrast, post-hoc explanation techniques focus on making sense of black-box models after they've been trained. These approaches don't change the model itself; instead, they create explanations for the outputs it produces. For instance, Ribeiro et al. (2016) introduced LIME, a popular model-agnostic method that approximates local decision boundaries to clarify individual predictions. Similarly, Lundberg and Lee (2017) came up with SHAP, a method grounded in cooperative game theory that consistently assigns importance values to features across different predictions.

Another significant contribution to the field of explainable AI (XAI) is the creation of detailed taxonomies and surveys that organize the expanding array of explainability methods. Guidotti et al. (2018) offer a thorough review of explanation strategies for black-box models, stressing the importance of balancing interpretability, fidelity, and usability. Additionally, Arrieta et al. (2020) point out that explainability should be seen as a crucial aspect of responsible AI, fostering trust and ensuring a human-centered approach to deployment.

The foundational concepts of XAI highlight the growing understanding that accuracy alone isn't enough in high-stakes situations. AI systems also need to be interpretable, transparent, and aligned with ethical and societal standards. Grasping these foundations is vital before diving into the specific methods, applications, and challenges that come with research in explainable AI.

International Journal of Engineering, Science and Humanities
An international peer reviewed, refereed, open-access journal
Impact Factor 8.3 www.ijesh.com ISSN: 2250-3552

3. Taxonomy of Explainability Methods in XAI

Explainable Artificial Intelligence, or XAI, covers a broad spectrum of techniques designed to enhance the transparency and interpretability of machine learning models. Given the variety of AI applications and the intricate nature of today's predictive systems, researchers have come up with several ways to categorize explainability approaches. These classifications often depend on how explanations are generated, the type of model being explained, and the level of interpretability offered.

One widely accepted classification splits XAI methods into two main categories: intrinsic interpretability and post-hoc explainability. Intrinsic methods refer to models that are naturally understandable, like linear regression, decision trees, and rule-based systems. These models let users see exactly how predictions are made, which is crucial for high-stakes decision-making where transparency is key (Rudin, 2019). However, these simpler models might not always deliver the same predictive accuracy as more complex black-box models.

On the other hand, post-hoc explanation methods focus on interpreting complex models after they've been trained. These techniques have gained traction because they allow the use of powerful algorithms, such as deep neural networks and ensemble models, while still offering insights that are understandable to humans. As noted by Guidotti et al. (2018), post-hoc explanations can be further broken down into several categories, including feature importance methods, surrogate models, example-based explanations, and visualization techniques.

Another important distinction is between model-agnostic and model-specific explanation methods. Model-agnostic approaches can be applied to any machine learning model, no matter its internal structure. For instance, LIME creates local explanations by approximating a black-box model with a simpler, interpretable model around a specific prediction (Ribeiro et al., 2016). Similarly, SHAP offers consistent feature attribution values based on Shapley game theory, which supports both local and global interpretability (Lundberg & Lee, 2017).

Model-specific methods are tailored for specific families of algorithms. Take saliency maps and gradient-based explanation techniques, for example; they're commonly used in deep learning models, especially in computer vision tasks (Simonyan et al., 2014). These methods leverage model gradients to pinpoint which input areas have the most significant impact on predictions.

Explanations can also be divided into local and global categories. Local explanations zoom in on individual predictions, which is particularly important in fields like healthcare or law, where every decision needs to be justified on a case-by-case basis. On the other hand, global

International Journal of Engineering, Science and Humanities
An international peer reviewed, refereed, open-access journal
Impact Factor 8.3 www.ijesh.com ISSN: 2250-3552

explanations provide insights into the overall behavior of the model, helping to foster a broader understanding and facilitate the auditing of decision-making systems (Molnar, 2022).

Arrieta et al. (2020) point out that for explanations to be truly effective, they need to combine technical transparency with a human-centered approach. This ensures that the explanations resonate with stakeholders, whether they are doctors, policymakers, or end users.

The classification of XAI methods underscores the increasing effort to organize interpretability research, paving the way for the development of reliable AI systems in high-stakes scenarios. Grasping these categories lays the groundwork for assessing specific techniques and determining their appropriateness in critical areas.

4. Explainability Techniques in High-Stakes Decision-Making

Explainability techniques are at the heart of Explainable Artificial Intelligence (XAI), helping users make sense of, trust, and validate machine learning predictions, especially in sensitive and high-stakes areas. In systems where decisions carry significant weight, having clear explanations is crucial—not just for transparency, but also for accountability, fairness, and meeting regulatory standards. Over the last ten years, a variety of explanation methods have emerged, with several becoming essential tools in applied XAI research.

One of the most popular techniques is local surrogate explanation models, with LIME (Local Interpretable Model-Agnostic Explanations) leading the way. Introduced by Ribeiro et al. in 2016, LIME takes a model-agnostic approach to explain individual predictions by simplifying the complex black-box model into a more interpretable one, focusing on the local area around the instance being predicted. This makes LIME particularly useful in fields like healthcare and finance, where justifying decisions on a case-by-case basis is often necessary.

Another key method is SHAP (Shapley Additive Explanations), created by Lundberg and Lee in 2017. SHAP uses principles from Shapley game theory to provide feature attribution values, delivering consistent and theoretically sound explanations. It's widely utilized for both local explanations (specific decisions) and global explanations (overall feature importance), making it a strong option for auditing AI models in high-stakes situations.

Beyond feature attribution methods, visual explanation techniques have also gained traction, especially in deep learning applications like medical imaging and autonomous driving. Simonyan et al. proposed saliency maps in 2014, which highlight the parts of an input image that most significantly impact the model's output. These visualization tools empower clinicians and researchers to discern whether AI systems are focusing on relevant medical features or getting sidetracked by misleading patterns.

Counterfactual explanations have become a key focus in the world of explainable AI (XAI). These counterfactuals illustrate how an input needs to change to shift the model's

International Journal of Engineering, Science and Humanities
An international peer reviewed, refereed, open-access journal
Impact Factor 8.3 www.ijesh.com ISSN: 2250-3552

decision, providing clear and actionable insights that are easy for people to understand. This kind of explanation is especially valuable in areas like loan approvals or parole decisions, where decision-makers often want to grasp the most significant factors that influenced an outcome (Molnar, 2022).

However, it's important to note that these explanation techniques do have their drawbacks. Research by Slack et al. (2020) showed that post-hoc methods like LIME and SHAP can be susceptible to adversarial attacks, leading to misleading explanations even when the predictions stay the same. This raises some serious questions about how reliable these explanation tools are in critical real-world situations.

Techniques for explainability—like LIME, SHAP, saliency maps, and counterfactual frameworks—are vital parts of trustworthy AI systems. Choosing the right methods really depends on the specific application, the type of model being used, and what stakeholders need. Ongoing research is essential to enhance the robustness, human interpretability, and practical usability of these explanation methods.

5. Applications of Explainable AI in High-Stakes Domains

Explainable Artificial Intelligence (XAI) has become increasingly important as it's being adopted in areas where decisions can have serious impacts on individuals and society. In high-stakes situations, AI systems need to do more than just deliver accurate predictions; they must also provide transparency, accountability, and build trust. XAI techniques are vital in helping human decision-makers by offering clear insights into how models operate. Key application areas for explainable AI include healthcare, finance, criminal justice, and autonomous systems.

5.1 Explainable AI in Healthcare

Healthcare is one of the most sensitive and impactful fields where AI systems are being used more and more for diagnosis, prognosis, and treatment recommendations. Deep learning models have proven effective in medical imaging and predicting diseases, but their black-box nature can hinder acceptance in clinical settings. It's crucial for medical professionals to understand and validate AI outputs before they can be applied in patient care.

Caruana et al. (2015) highlighted the significance of interpretable models in predicting pneumonia risk, showing that clear explanations can help avoid harmful decision-making mistakes. Likewise, Tonekaboni et al. (2019) pointed out that clinicians need context-aware and actionable explanations, rather than just technical feature importance scores. By enhancing trust, explainable AI in healthcare supports clinical decision-making and allows for the ethical use of AI technologies.

5.2 Explainable AI in Finance

International Journal of Engineering, Science and Humanities
An international peer reviewed, refereed, open-access journal
Impact Factor 8.3 www.ijesh.com ISSN: 2250-3552

In the finance sector, decision-making systems are increasingly turning to machine learning for tasks like credit scoring, fraud detection, bankruptcy prediction, and risk management. These decisions can greatly influence individuals' access to loans, job opportunities, and overall financial stability. That's why transparency and fairness are so essential.

Feng et al. (2021) pointed out how crucial explainable AI is in the financial sector, emphasizing that being able to interpret AI decisions helps institutions meet regulatory requirements and minimizes biased outcomes. Similarly, Barboza et al. (2017) showcased how effective machine learning models can be in predicting bankruptcies, where having clarity can lead to a better grasp of financial risk factors. Explainable AI ensures that automated financial choices are transparent, equitable, and socially responsible.

5.3 Explainable AI in Criminal Justice

AI technologies are making their way into the criminal justice system for tasks like predictive policing, assessing the risk of reoffending, and supporting sentencing decisions. However, these uses bring up significant ethical issues, particularly concerning bias and discrimination. A notable case is COMPAS, which faced criticism for its algorithmic decisions leading to racial disparities (Angwin et al., 2016).

Rudin (2019) contends that we should steer clear of black-box models in such critical situations, advocating for interpretable models that promote accountability. Explainability is essential in legal frameworks to ensure fairness, protect human rights, and allow for decisions to be questioned or justified.

5.4 Explainable AI in Autonomous and Safety-Critical Systems

Autonomous technologies, including self-driving cars, robotics, and industrial automation, are increasingly reliant on AI for decision-making. In these high-stakes environments, failures can lead to accidents and even fatalities. Explainability is key for engineers and regulators to comprehend how these systems operate, enhance reliability, and develop safer autonomous solutions.

Amodei et al. (2016) highlighted significant safety challenges in AI systems, stressing the importance of transparency to avoid unintended harmful actions. Explainable frameworks can improve monitoring and control in complex autonomous settings.

6. Challenges and Limitations of Explainable AI in High-Stakes Decision-Making

Even though we've made great strides in Explainable Artificial Intelligence (XAI), putting these systems to work in high-stakes areas still comes with its fair share of challenges and limitations. While the goal of XAI methods is to boost transparency and trust, a range of technical, practical, and ethical issues can really get in the way of their effectiveness.

International Journal of Engineering, Science and Humanities
An international peer reviewed, refereed, open-access journal
Impact Factor 8.3 www.ijesh.com ISSN: 2250-3552

6.1 The Balancing Act between Accuracy and Interpretability

One ongoing challenge is finding the right balance between how accurate a model is and how interpretable it is. Models that are easy to understand, like linear regression or decision trees, provide straightforward explanations but might not perform as well as more complex black-box models, such as deep neural networks or ensemble methods (Rudin, 2019). On the flip side, post-hoc explanations can shed light on these complex models, but they might not fully grasp the model's inner workings, which can lead to incomplete or even misleading insights (Slack et al., 2020). Figuring out how to balance performance with explainability is still a key question in research.

6.2 The Subjective Nature of Explanation Evaluation

Assessing the quality of explanations is inherently subjective. What one person sees as a "useful" explanation can differ based on the stakeholder, context, and domain (Doshi-Velez & Kim, 2017). For instance, clinicians might look for explanations that lead to actionable insights, while regulators may be more concerned with transparency and compliance. This variability makes it tough to create standardized evaluation metrics and complicates the ability to compare different XAI methods.

6.3 Scalability and Real-Time Constraints

In high-stakes scenarios like self-driving cars, industrial automation, or real-time financial trading, there's a pressing need for explanations to be generated swiftly and at scale. Unfortunately, current post-hoc methods like LIME and SHAP can be quite resource-heavy, which makes them less practical for real-time use (Molnar, 2022). This creates a bit of a dilemma: how do we provide detailed and accurate explanations while still keeping the system running smoothly, especially in fast-paced environments where quick decisions are crucial?

6.4 Risk of Misleading or Adversarial Explanations

One of the downsides of post-hoc explanation methods is their vulnerability to producing misleading or inconsistent results, particularly when dealing with complex or poorly understood models. Research by Slack et al. (2020) showed that LIME and SHAP can be manipulated in ways that yield explanations that seem plausible but don't actually represent the model's true decision-making process. This is especially worrisome in high-stakes fields, where stakeholders depend on these explanations for accountability and trust.

6.5 Ethical, Legal, and Regulatory Challenges

When deploying XAI in sensitive areas, there are also ethical, legal, and regulatory hurdles to consider. Take GDPR, for instance, which requires transparency in automated decision-making but doesn't specify how to achieve that explainability, leaving organizations in a bit of a bind regarding compliance (European Union, 2016). Ethical issues can arise, such as

International Journal of Engineering, Science and Humanities
An international peer reviewed, refereed, open-access journal
Impact Factor 8.3 www.ijesh.com ISSN: 2250-3552

biased explanations, a lack of inclusivity in design, and the potential for AI decisions to be misused, highlighting the need for solid governance frameworks and specific guidelines tailored to different domains.

6.6 Human-Centered Interpretability

The gap between technical explanations and human comprehension is one of our main problems. Even though the explanations generated by XAI algorithms may be theoretically valid, end users or domain experts may find it difficult to understand them in a way that makes sense (Tonekaboni et al., 2019). Applying human-centered design principles is crucial because it ensures that the explanations we offer genuinely meet the contextual, professional, and cognitive demands of all parties concerned.

Despite significant advancements in XAI approaches, we continue to face challenges such as the trade-offs between interpretability and performance, subjectivity, scaling problems, adversarial attack vulnerability, and the requirement for ethical governance. In order to develop AI systems that are trustworthy, transparent, and accountable in high-stakes decision-making situations, it is imperative to address these issues.

7. Future Directions and Research Opportunities in Explainable AI

As Explainable Artificial Intelligence (XAI) keeps evolving, a number of exciting research paths are opening up to tackle existing challenges and improve how AI is used in critical decision-making scenarios. One key area of focus is the creation of standardized, domain-specific metrics to evaluate the quality of explanations. Right now, assessing XAI methods tends to be quite subjective, which makes it tough to compare different approaches or gauge their effectiveness across various applications (Doshi-Velez & Kim, 2017). By establishing metrics that take into account fidelity, completeness, usability, and human trust, we can lay a stronger groundwork for validating explanation techniques.

Another important aspect is designing explanations that are human-centered and aware of their context. Just because an explanation is technically correct doesn't mean it meets the needs or cognitive models of the end-users. Context-aware and actionable explanations, especially in fields like healthcare, finance, and legal systems, can enhance decision-making and build greater trust among stakeholders (Tonekaboni et al., 2019). New research is highlighting the importance of interactive and adaptive explanation systems that can customize insights based on the user's expertise and the specific context of their decision-making.

Additionally, there's a growing interest in integrating causal and counterfactual reasoning with existing XAI methods. Counterfactual explanations, which show the minimal changes needed to alter a model's outcome, offer actionable insights that are particularly useful in scenarios like loan approvals or clinical decision support (Molnar, 2022). Merging causal

International Journal of Engineering, Science and Humanities
An international peer reviewed, refereed, open-access journal
Impact Factor 8.3 www.ijesh.com ISSN: 2250-3552

inference with traditional interpretability approaches could significantly boost both the reliability and practical utility of explanations in complex, high-stakes environments.

Making sure that explanations are both robust and secure is a significant challenge we face. Post-hoc explanation methods can easily fall prey to adversarial manipulation or noise, leading to misleading explanations, even when the model's predictions are spot on (Slack et al., 2020). It's crucial to develop methods that can withstand these vulnerabilities, especially in areas where human lives, financial security, or legal outcomes hang in the balance.

Ethical, legal, and governance issues are at the forefront of responsibly adopting explainable AI (XAI). It's vital to create clear guidelines for accountability, fairness, and transparency that are tailored to specific sectors (European Union, 2016). To build strong governance frameworks, we need collaboration across disciplines—AI researchers, ethicists, policymakers, and domain experts must work together to ensure that AI systems are not only technically sound but also socially responsible.

Scalability and real-time explainability are crucial areas ripe for innovation. Many high-stakes applications, like autonomous vehicles and financial trading, demand that explanations be generated quickly and at scale. By optimizing post-hoc methods and utilizing model-specific techniques, we can cut down on computational overhead, making it feasible to deploy these systems in fast-paced, time-sensitive environments. Ultimately, future research in XAI should focus on creating explanation methods that are robust, interpretable, human-centered, and ethically aligned, paving the way for trustworthy and accountable AI in critical decision-making scenarios.

8. Conclusion

Explainable Artificial Intelligence (XAI) has become a crucial part of implementing AI systems in areas where decisions carry significant weight, such as healthcare and finance. In these fields, transparency, accountability, and ethical responsibility are paramount. This review highlights that while advanced machine learning models can deliver impressive predictive results, their black-box nature creates hurdles in terms of understanding, trust, and meeting regulatory standards. XAI offers a variety of methods—like inherently interpretable models, post-hoc techniques such as LIME and SHAP, saliency-based visualizations, and counterfactual explanations—to tackle these issues. The use of XAI in sectors like healthcare, finance, criminal justice, and autonomous systems showcases its ability to improve decision-making, enhance accountability, identify bias, and build user trust. Nevertheless, challenges remain, such as the balance between accuracy and interpretability, the subjective nature of evaluating explanations, scalability concerns, susceptibility to adversarial attacks, and ethical or legal dilemmas. Moving forward, XAI research should prioritize creating standardized evaluation metrics, developing

human-centered and context-aware explanations, ensuring robust and secure methods, and finding scalable solutions that comply with ethical and regulatory guidelines. By overcoming these obstacles, XAI can help create AI systems that are not only precise but also transparent, trustworthy, and aligned with human values, ultimately encouraging responsible and socially beneficial use of AI in critical environments.

References

1. Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., & Mané, D. (2016). Concrete problems in AI safety. *arXiv preprint arXiv:1606.06565*.
2. Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016). Machine bias. *ProPublica.* https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing
3. Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., … Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion, 58*, 82–115. https://doi.org/10.1016/j.inffus.2019.12.012
4. Barboza, F., Kimura, H., & Altman, E. (2017). Machine learning models and bankruptcy prediction. *Expert Systems with Applications, 83*, 405–417. https://doi.org/10.1016/j.eswa.2017.04.006
5. Caruana, R., Lou, Y., Gehrke, J., Koch, P., Sturm, M., & Elhadad, N. (2015). Intelligible models for healthcare: Predicting pneumonia risk and hospital readmission. *Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1721–1730. https://doi.org/10.1145/2783258.2788613
6. Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*.
7. European Union. (2016). General Data Protection Regulation (GDPR). *Official Journal of the European Union*.
8. Feng, X., Cai, Z., Li, X., & Li, Y. (2021). Explainable AI in finance: Applications and challenges. *Journal of Risk and Financial Management, 14*(9), 430. https://doi.org/10.3390/jrfm14090430
9. 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), 1–42. https://doi.org/10.1145/3236009
10. Lipton, Z. C. (2018). The mythos of model interpretability. *Communications of the ACM, 61*(10), 36–43. https://doi.org/10.1145/3233231
11. Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems, 30*, 4765–4774.

12. Molnar, C. (2022). *Interpretable machine learning* (2nd ed.). Lulu Press.
13. National Institute of Standards and Technology. (2023). Artificial Intelligence Risk Management Framework (AI RMF 1.0). U.S. Department of Commerce.
14. 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. https://doi.org/10.1145/2939672.2939778
15. Rudin, C. (2019). Stop explaining black box machine learning models for high-stakes decisions and use interpretable models instead. *Nature Machine Intelligence, 1*, 206–215. https://doi.org/10.1038/s42256-019-0048-x
16. Simonyan, K., Vedaldi, A., & Zisserman, A. (2014). Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*.
17. 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 Conference on Artificial Intelligence*, 34(04), 1801–1808. https://doi.org/10.1609/aaai.v34i04.6046
18. Tonekaboni, S., Joshi, S., McCradden, M. D., & Goldenberg, A. (2019). What clinicians want: Contextualizing explainable machine learning for clinical end use. *arXiv preprint arXiv:1905.05134*.