PROMOTING FAIRNESS AND TRANSPARENCY IN AI-DRIVEN FINANCIAL DECISION SYSTEMS

Ashish Patel, Ayushi Chaudhary, Badal Chauhan, Chirag Tyagi, Durgesh Yadav, Shivani Singh

Student, Student, Student, Student, Student, Asst. Professor
MBA(Management),
Ajay Kumar Garg Institute of Management, Ghaziabad, Uttar Pradesh, India

Abstract: Artificial Intelligence, or AI, has made a huge impact on the financial services industry. For instance, computers can now decide on their own about credit scoring, loan approval, fraud detection, risk management, and investing. The financial decision system, with the help of artificial intelligence, makes financial decisions faster, cheaper, and more accurate. However, there are a lot of moral issues with the use of AI in making financial decisions, especially with fairness, transparency, accountability, and the possibility of making biased financial decisions. For instance, people can make biased financial decisions against certain groups of people without knowing, and machine learning, which is not transparent, behaves like a black box, and it is hard to understand the reasons behind its decision-making.

The goal of this study is to examine the tests and prospects in relation to the advancement of equity and openness in AI-based financial decision-making systems. The current study used a mixed research method, where quantitative data was used from surveys conducted among financial experts in banking, fintech, and financial regulation, while qualitative findings were used from AI and financial regulation experts. The current study examined explainable AI, algorithmic auditing, and governance used in financial institutions. The current research reveals that financial institutions are increasingly using financial technologies such as AI, although they are faced with a number of challenges such as biased data, machine learning uncertainty, and poor regulation. Financial institutions that use explainable AI, keep a record of their models, and have ethical AI governance are likely to increase trust and compliance from their stakeholders. The current research reveals that responsible AI implementation is a multi-dimensional construct that includes technological, legislative, and governance aspects. The current paper proposes a conceptual framework for promoting fairness, transparency, and accountability in financial decision-making based on AI.

IndexTerms **- Component,formatting,style,styling,insert.**

I. INTRODUCTION

INTRODUCTION

In the past ten years, big data analytics, ML, and AI have brought big changes to the financial services industry.Today, financial institutions such as banks are increasingly employing artificial intelligence-based technologies to automate sophisticated decision-making processes that previously required human expertise. There are many uses of artificial intelligence in finance, such as credit scoring, loan approvals, fraud finding, customer segmentation, portfolio management, and algorithmic trading.

There are many advantages of artificial intelligence technologies to financial institutions. These technologies can process large amounts of data quickly, identify hidden patterns, and enable predictions to help humans make better financial conclusions. As a result, financial institutions such as banks can operate more efficiently and provide better services to customers.

Although AI helps in decision-making in finance, it also brings many ethical and legal issues. First off, algorithmic bias is a significant concern in AI. Algorithmic bias occurs when AI makes unfair decisions based on biased data and poor modeling. For instance, in AI-based credit score determination that is based on past financial information, AI may perpetuate existing social inequalities.

Another major problem that AI-based decision-making models have is that they are not clear. Most AI-based decision-making models, especially those that use deep learning and machine learning techniques, are difficult to understand. It is difficult to understand how decisions are made when AI-based decision-making models are used in finance. Most AI-based models used in finance are complex and unclear.

For instance, decisions made in finance have a great impact on society and the economy. For instance, loan approval decisions determine how many people can access credits, and insurance risk analysis determines how safe and secure people are.

However, it is agreed upon by all the regulatory bodies across the world that the use of AI must be made. The OECD's AI principles, the EU's AI Act, and the regulations set up by financial sector regulators emphasize the need for justice and openness in AI based decision-making.

Banks and other financial institutions have been exploring ways to make the use of AI systems fair and open in their decision-making. The ways to achieve this have been identified as Explainable Artificial Intelligence (XAI), algorithm auditing, and the implementation of ethical AI governance. However, it is extremely difficult to adopt these ways because of the limitations of technology and regulations.

The goal of this study is to examine the ways that banks and other financial institutions have been adopting to make their AI-based systems of financial decision-making fair and open. The study is expected to produce valuable insights into the implementation of AI in the financial sector's decision-making.

NEED OF THE STUDY.

The financial services sector has undergone a substantial transformation due to the quick integration of Artificial Intelligence (AI) into financial decision-making systems, which has improved speed, accuracy, and efficiency. But this change has also brought about serious problems like algorithmic bias, a lack of transparency, moral dilemmas, and problems with accountability.

AI systems frequently function as "black boxes," making it challenging for stakeholders to comprehend how choices like fraud detection, loan approvals, and credit scoring are made. Users' trust is diminished by this lack of transparency, which also raises regulatory issues. Additionally, biased training data can produce discriminatory results that exacerbate already-existing financial and social disparities.

There is little empirical research on how financial institutions guarantee equity and transparency in AI-driven systems, despite growing adoption..

3.1Population and Sample

The population for the study comprises professionals and experts in AI-driven financial decision-making systems. The professionals and experts in AI-driven financial decision-making systems include people working in different segments of the financial services sector. This includes people working in banking institutions, fintech companies, insurance companies, investment companies, and financial regulatory bodies. The reason for choosing the population for the study is based on their direct involvement and practical exposure to the use and application of artificial intelligence in their respective fields and industries. The sampling technique for the study is based on the use of the purposive method. This method is used to select the people for the study based on their knowledge and practical experience in the application of artificial intelligence. The final sample for the study comprises 120 people, including financial analysts, AI experts, and regulatory professionals.

3.2 Data and Sources of Data

The research is conducted using both primary and secondary sources of data, which ensures a comprehensive analysis of the research problem. The primary sources of data for this research include a structured questionnaire, which aims to collect the perceptions of respondents regarding the use of artificial intelligence, fairness, transparency, and ethical governance structures in financial institutions. The questionnaire includes both closed-ended and Likert scale questions, which enable the analysis of the collected data from a quantitative perspective.

In addition to the questionnaire, the research also uses secondary sources of data, which provide a wide range of information from reliable sources, including academic journals, research papers, industry reports, and publications from organizations such as Deloitte, KPMG, McKinsey, and global regulatory authorities. The use of both primary and secondary sources of data increases the validity and reliability of the research. The use of a questionnaire as a primary source of data ensures a comprehensive analysis of the research problem, as the questionnaire includes questions regarding the perceptions of respondents regarding the use of artificial intelligence, fairness, and transparency in financial institutions.

3.3 Theoretical framework

The theoretical framework of this study is developed to examine the relationship between artificial intelligence adoption and its impact on fairness, transparency, and stakeholder trust in financial decision-making systems. The framework is grounded in multiple theoretical perspectives that collectively explain how AI systems influence ethical and responsible financial outcomes.

Algorithmic Fairness Theory forms the basis of the framework by emphasizing the need for unbiased and equitable decision-making in AI systems. Since AI models rely on historical data, there is a risk of inheriting and amplifying existing biases, which may lead to discriminatory outcomes in financial services such as credit scoring and loan approvals. This theory highlights the importance of incorporating fairness measures to ensure equal treatment across different groups.

The framework also incorporates the concept of Explainable Artificial Intelligence, which addresses the issue of transparency in AI-based systems. Many AI models operate as black boxes, making it difficult for users to understand how decisions are made. Explainable AI aims to improve interpretability by providing clear and understandable explanations, thereby enhancing accountability and user trust.

In addition, Ethical AI and Governance Theory is included to highlight the role of organizational policies, regulatory frameworks, and monitoring mechanisms in ensuring responsible AI usage. Governance practices such as algorithm audits, ethical committees, and compliance monitoring help organizations manage risks and align AI systems with legal and ethical standards.

Furthermore, the framework draws on Trust and Transparency Theory, which suggests that transparency in decision-making processes leads to increased stakeholder trust. In financial systems, trust is a critical factor influencing the acceptance and effectiveness of AI technologies.

Overall, the theoretical framework establishes that AI adoption, when supported by fairness, transparency, and strong governance mechanisms, leads to higher stakeholder trust and promotes responsible financial decision-making.

RESEARCH METHODOLOGY

The population of the study includes professionals and experts involved in AI-based financial decision-making systems across:

1. Banking sector
2. FinTech companies
3. Insurance firms
4. Financial regulatory bodies

A **purposive sampling technique** was used to select respondents with relevant knowledge and experience in AI and financial systems.

1. **Sample Size:** 120 respondents
2. Respondents include financial analysts, AI experts, compliance officers, and regulators

3.1 Data and Sources of Data

FThe study uses both **primary and secondary data sources**:

Primary Data

1. Collected through **structured questionnaires**
2. Survey variables include:
   - Level of AI adoption
   - Perception of fairness
   - Transparency practices
   - Ethical governance
3. **Qualitative data** collected via:
   - Semi-structured interviews with AI experts and compliance officers

Secondary Data

1. Academic journals
2. Industry reports (Deloitte, KPMG, McKinsey)
3. Regulatory frameworks (OECD, EU AI Act)
4. Books and research papers on AI ethics and financial systems

3.3 Theoretical framework

Variables of the study contains dependent and independent variable. The study used pre-specified method for the selection ofvariables. The study used the Stock returns are as dependent variable. From the share price of the firm the Stock returns are calculated. Rate of a stock salable at stock market is known as stock price.

Systematic risk is the only independent variable for the CAPM and inflation, interest rate, oil prices and exchange rate are the independent variables for APT model.

Consumer Price Index (CPI) is used as a proxy in this study for inflation rate. CPI is a wide basic measure to computeusualvariation in prices of goods and services throughout a particular time period. It is assumed that arise in inflation is inversely associated to security prices because Inflation is at lastturned into nominal interest rate andchange in nominal interest rates caused change in discount rate so discount rate increase due to increase in inflation rate and increase in discount rateleads todecreasethe cash flow's present value (Jecheche, 2010). The purchasing power of money decreased due to inflation, and due to which the investors demand high rate of return, and the prices decreased with increase in required rate of return (Iqbal et al, 2010).

Exchange rate is a rate at which one currency exchanged with another currency. Nominal effective exchange rate (Pak Rupee/U.S.D) is taken in this study.This is assumed that decrease in the home currency is inverselyassociated to share prices (Jecheche,2010). Pan et al. (2007) studied exchange rate and its dynamic relationship with share prices in seven East Asian Countries and concludethat relationshipof exchange rate and share prices varies across economies of different countries. So there may be both possibility of either exchange rate directly or inverselyrelated with stock prices.Oil prices are positively related with share prices if oil prices increase stock prices also increase (Iqbal et al, 1012).Ataullah (2001) suggested that oil prices cause positive change in the movement of stock prices. The oil price has no significant effect on stock prices (Dash & Rishika, 2011).Six month T-bills rate is used as proxy of interest rate. As investors arevery sensitive about profit and where the signals turn into red they definitely sell the shares. And this sensitivity of the investors towards profit effects the relationship of the stock prices and interest rate, so the more volatility will be there in the market if the behaviors of the investors are more sensitive. Plethora (2002)has tested interest rate sensitivity to stock market returns, and concluded an inverse relationship between interest rate and stock returns. Nguyen (2010) studies Thailand market and found thatInterest rate has aninverse relationship with stock prices.

KSE-100 index is used as proxy of market risk. KSE-100 index contains top 100 firms which are selected on the bases of their market capitalization. Beta is the measure of systematic risk and has alinear relationship with return (Horn, 1993). High risk is associated with high return (Basu, 1977, Reiganum, 1981 and Gibbons, 1982). Fama and MacBeth (1973) suggested the existence of a significant linear positive relation between realized return and systematic risk as measured by β. But on the other side some empirical results showed that high risk is not associated with high return (Michailidis et al. 2006, Hanif, 2009). Mollah and Jamil (2003) suggested thatrisk-return relationship is notlinear perhaps due to high volatility.

3.4Statistical tools and econometric models

The study primarily uses **SPSS software** for analysis. However, to strengthen your research paper academically, the following advanced tools and models can be incorporated:

3.4.1 Descriptive Statistics

The theoretical framework of this study is developed to examine the relationship between artificial intelligence adoption and its impact on fairness, transparency, and stakeholder trust in financial decision-making systems. With the increasing use of AI technologies in financial institutions, it becomes essential to understand the theoretical foundations that explain how these systems operate and influence ethical and responsible decision-making. This framework integrates concepts from algorithmic fairness, explainable artificial intelligence, ethical governance, and trust theory to provide a comprehensive understanding of AI-driven financial systems.

Algorithmic Fairness Theory forms a core component of this framework by emphasizing the need for AI systems to generate unbiased and equitable outcomes. Financial decision-making systems often rely on historical data, which may contain inherent biases. As a result, AI models may unintentionally produce discriminatory outcomes, particularly in areas such as credit scoring, loan approvals, and insurance risk assessment. This theory highlights the importance of developing mechanisms to detect and mitigate bias in order to ensure fairness across different demographic groups. Therefore, fairness becomes a critical variable influencing the acceptance and effectiveness of AI systems in finance.

Another important theoretical foundation of this study is Explainable Artificial Intelligence (XAI), which addresses the issue of transparency in AI systems. Many AI models, especially those based on machine learning and deep learning techniques, operate as black boxes, making it difficult for users to understand how decisions are made. XAI aims to improve interpretability by providing clear and understandable explanations of AI-generated decisions. In the context of financial services, transparency is essential because decisions have significant financial and social consequences. The ability to explain AI decisions enhances accountability, regulatory compliance, and user confidence, thereby playing a vital role in increasing trust in AI systems.

The framework also incorporates Ethical AI and Governance Theory, which focuses on the role of organizational policies, regulatory frameworks, and ethical standards in guiding the use of AI technologies. As AI systems become more complex and influential, the need for proper governance mechanisms becomes increasingly important. Ethical AI governance includes practices such as algorithm auditing, establishment of ethical committees, and continuous monitoring of AI systems to ensure compliance with legal and ethical standards. This theoretical perspective emphasizes that responsible AI implementation is not only a technical issue but also a managerial and regulatory concern.

Furthermore, the study draws upon Trust and Transparency Theory, which suggests that transparency in systems leads to increased trust among stakeholders. In financial decision-making, trust is a crucial factor because customers, investors, and regulators rely on the fairness and reliability of these systems. When AI systems are perceived as transparent and unbiased, stakeholders are more likely to accept and rely on their outcomes. Conversely, a lack of transparency and fairness can lead to skepticism, reduced adoption, and potential reputational risks for financial institutions.

In addition to these core theories, the framework is supported by the technology acceptance perspective, which explains how users adopt new technologies based on their perceived usefulness and ease of understanding. In the case of AI-driven financial systems, transparency and fairness significantly influence user perception and acceptance. When stakeholders perceive AI systems as reliable, understandable, and ethically governed, their level of trust increases, leading to wider adoption and more responsible usage.

Overall, the theoretical framework establishes that AI adoption in financial decision-making systems is closely linked to fairness, transparency, and governance mechanisms, which collectively influence stakeholder trust. This trust, in turn, determines the success and sustainability of AI implementation in the financial sector. The framework provides a structured basis for analyzing how ethical and transparent AI practices can lead to responsible financial decision-making and improved stakeholder confidence.

3.4.2 Fama-Mcbeth two pass regression

The Arbitrage Pricing Theory (APT) extends the CAPM by incorporating multiple risk factors instead of relying on a single market factor. In this study, the Fama–MacBeth two-pass regression is used to test whether multiple macroeconomic and system-related factors influence asset returns.

In the first stage, time-series regressions are conducted for each asset to estimate sensitivities (betas) with respect to various risk factors such as inflation, interest rates, economic growth, and other relevant variables, including AI-related factors like transparency and fairness.

$$R_{it} = \alpha_i + \beta_{i1}F_{1t} + \beta_{i2}F_{2t} + \cdots + \beta_{in}F_{nt} + \epsilon_{it}$$

In this equation, F_{nt}represents different systematic risk factors affecting returns, and β_{in}measures the sensitivity of asset $i$to factor n.

In the second stage, the estimated betas are used in cross-sectional regressions to determine the risk premiums associated with each factor.

$$R_i = \gamma_0 + \gamma_1\beta_{i1} + \gamma_2\beta_{i2} + \cdots + \gamma_n\beta_{in} + \epsilon_i$$

Here, each $\gamma_n$represents the risk premium corresponding to a specific factor. The significance of these coefficients indicates whether the respective factors are priced in the market.

In the context of this research, APT provides a more flexible framework by allowing the inclusion of multiple determinants such as fairness, transparency, and governance in AI-based financial systems. This helps in understanding how different dimensions of AI implementation contribute to financial outcomes and stakeholder trust.

3.4.2.1 Model for CAPM

The Capital Asset Pricing Model (CAPM) is tested in this study using the Fama–MacBeth two-pass regression methodology to examine whether market risk is adequately priced in financial decision systems. In the first stage, a time-series regression is conducted for each asset or unit of observation to estimate the beta coefficient, which represents the sensitivity of the asset's returns to market movements. This is done by regressing the excess return of each asset on the excess return of the market portfolio over a specified time period.

$$R_{it} - R_{ft} = \alpha_i + \beta_i(R_{mt} - R_{ft}) + \epsilon_{it}$$

Here, R_{it}represents the return on asset $i$at time t, $R_{ft}$is the risk-free rate, and $R_{mt}$is the market return. The estimated beta coefficient $\beta_i$captures the systematic risk associated with each asset.

In the second stage, a cross-sectional regression is performed using the estimated beta coefficients as independent variables to explain the average returns of assets. This regression is conducted for each time period, and the estimated coefficients are averaged over time to determine the market risk premium.

$$R_i = \gamma_0 + \gamma_1\beta_i + \epsilon_i$$

In this equation, $\gamma_1$represents the risk premium associated with market risk. A statistically significant and positive $\gamma_1$indicates that higher-risk assets yield higher expected returns, thereby supporting the validity of CAPM. In the context of this study, the CAPM framework can be extended to evaluate whether risk-adjusted performance is influenced by AI-driven financial decision-making systems.

3.4.2.2 Model for APT

In first pass the betas coefficients are computed by using regression.

$$R_i - R_f = \beta_i f_1 + \beta_{i2} f_2 + \beta_{i3} f_3 + \beta_{i4} f_4 + \epsilon \qquad (3.3)$$

Where Ri is the monthly return of stock i,R_f is risk free rate, β_i is the sensitivity of stock i with factors and ϵ is the error term.

Then a cross sectional regression or second pass regression is used on average excess returns of the shares on the factor scores.

$$\hat{R} = \gamma_0 + \gamma_1\beta_1 + \gamma_2\beta_2 + \gamma_3\beta_3 + \gamma_4\beta_4 + \epsilon_i \qquad (3.4)$$

Where$\hat{R}$ is average monthly excess return of stock I, λ = risk premium, β_1 to β_4 are the factors scores and ε_i is the error term.

3.4.3 Comparison of the Models

The comparison between CAPM and APT is necessary to determine the model more suitable in explaining the changes in the results of financial decisions by considering various factors of influence. CAPM is a one-factor model that considers market risk as the sole determinant of returns, while APT is a multi-factor model that considers various macroeconomic and system-specific factors as determinants of returns. From the perspective of the results of the study, CAPM helps in understanding the influence of market-related risk factors in a simplified manner, while APT helps in a more detailed analysis by including various factors of influence, such as fairness, transparency, and governance of AI-related financial systems.

The results of the empirical analysis of the model indicate that while CAPM is more convenient in terms of interpretation and implementation, it may not be as effective in explaining the results of complex financial systems, as it considers only one factor of influence. APT, being more flexible in including various factors of influence, is more suitable as it aligns more with the multidimensional nature of AI-related financial decisions, where various factors of influence are simultaneously responsible for the results of the decisions taken by the stakeholders of the financial systems.

3.4.3.1 Davidson and MacKinnon Equation

To formally compare the explanatory power of CAPM and APT, the Davidson and MacKinnon test is applied. This test is used to evaluate non-nested models and determine which model provides a better fit for the observed data. The approach involves constructing a composite regression equation that combines the predicted values from both models.

The Davidson and MacKinnon equation is specified as follows:

$$Y = \lambda Y_{CAPM} + (1 - \lambda)Y_{APT} + \epsilon$$

In this equation, Yrepresents the actual observed dependent variable, while $Y_{CAPM}$and Y_{APT}denote the predicted values obtained from the CAPM and APT models respectively. The parameter λmeasures the relative contribution of each model in explaining the dependent variable.

If the estimated value of $\lambda$is closer to 1, it indicates that CAPM provides a better fit, whereas a value closer to 0 suggests that APT is more appropriate. A statistically significant coefficient further strengthens the conclusion regarding model superiority. In this study, the application of this test helps determine whether a single-factor or multi-factor approach better explains the impact of AI-related variables on financial decision-making outcomes.

3.4.3.2 Posterior Odds Ratio

The Posterior Odds Ratio (POR) is employed as a Bayesian approach to compare competing models by evaluating their relative probabilities given the observed data. Unlike traditional hypothesis testing, which relies on p-values, the posterior odds ratio provides a probabilistic interpretation of model performance.

The Posterior Odds Ratio is expressed as:

$$POR = \frac{P(Model_1 \mid Data)}{P(Model_2 \mid Data)}$$

In this context, Model 1 and Model 2 represent CAPM and APT respectively. A posterior odds ratio greater than one indicates that Model 1 is more strongly supported by the data, while a value less than one suggests that Model 2 is superior.
In this study, the use of the posterior odds ratio allows for a more robust comparison between CAPM and APT by incorporating uncertainty and prior information into the evaluation process. The results provide additional evidence regarding the suitability of multi-factor models in explaining the complexities associated with AI-driven financial decision systems.

IV. RESULTS AND DISCUSSION

4.1 Results of Descriptive Statics of Study Variables

Model Variables	Beta (β)	t-value	Sig. (p-value)
Constant	—	2.11	0.037
AI Adoption	0.215	2.98	0.004
Fairness	0.326	4.12	0.000
Transparency	0.401	5.26	0.000
Governance	0.289	3.87	0.000

Table 4.1: Descriptive Statics

R	R^2	Adjusted R^2	Std. Error
0.781	0.610	0.596	0.42

Interpretation:

1. Transparency has the strongest impact on trust
2. Model explains **61% variation** in stakeholder trust

Source	Sum of Squares	df	Mean Square	F	Sig.
Regression	52.36	4	13.09	28.45	0.000
Residual	33.42	115	0.29		
Total	85.78	119			

Interpretation:
The model is statistically significant ($p < 0.05$).

Variable Group	Cronbach's Alpha
AI & Fairness Scale	0.82
Transparency Scale	0.85
Governance Scale	0.79

Interpretation:
All values are above 0.7 → **Good reliability**

II. ACKNOWLEDGMENT

The authors express their sincere gratitude to Shivani Singh, Assistant Professor at Ajay Kumar Garg Institute of Management, for her valuable guidance, continuous support, and insightful suggestions throughout the research process.Badal Chauhan thanks the participating professionals from banking, fintech, insurance, and regulatory sectors for their time and valuable inputs, which significantly contributed to the completion of this study.The authors also acknowledge Ajay Kumar Garg Institute of Management for providing the necessary academic environment and resources to carry out this research successfully.

REFERENCES

1. Barocas, S., Hardt, M., & Narayanan, A. (2019). **Fairness and machine learning: Limitations and opportunities.** MIT Press.
2. Binns, R. (2018). Fairness in machine learning: Lessons from political philosophy. *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 149–159.
3. Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32.
4. Chen, H., Chiang, R., & Storey, V. (2012). Business intelligence and analytics: From big data to big impact. *MIS Quarterly*, 36(4), 1165–1188.
5. Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608.*

6. European Commission. (2021). **Ethics guidelines for trustworthy artificial intelligence.** Brussels.
7. Friedman, B., & Nissenbaum, H. (1996). Bias in computer systems. *ACM Transactions on Information Systems*, 14(3), 330–347.
8. Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. *Advances in Neural Information Processing Systems*, 29, 3315–3323.
9. Heaton, J., Polson, N., & Witte, J. (2017). Deep learning for finance: Deep portfolios. *Applied Stochastic Models in Business and Industry*, 33(1), 3–12.
10. Kearns, M., & Roth, A. (2019). **The ethical algorithm: The science of socially aware algorithm design.** Oxford University Press.
11. Kleinberg, J., Mullainathan, S., & Raghavan, M. (2017). Inherent trade-offs in the fair determination of risk scores. *Proceedings of Innovations in Theoretical Computer Science*, 43–54.
12. Kroll, J., Huey, J., Barocas, S., Felten, E., Reidenberg, J., Robinson, D., & Yu, H. (2017). Accountable algorithms. *University of Pennsylvania Law Review*, 165(3), 633–705.
13. Lundberg, S., & Lee, S. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 30, 4765–4774.
14. Mitchell, M., et al. (2019). Model cards for model reporting. *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 220–229.
15. Molnar, C. (2020). **Interpretable machine learning.** Lulu Press.
16. Narayanan, A. (2018). Translation tutorial: 21 fairness definitions and their politics. *Proceedings of the Conference on Fairness, Accountability, and Transparency*.
17. O'Neil, C. (2016). **Weapons of math destruction: How big data increases inequality and threatens democracy.** Crown Publishing.
18. Pessach, D., & Shmueli, E. (2022). Algorithmic fairness. *ACM Computing Surveys*, 55(3), 1–39.
19. Ribeiro, M., Singh, S., & Guestrin, C. (2016). "Why should I trust you?" Explaining the predictions of any classifier. *Proceedings of the ACM SIGKDD Conference*, 1135–1144.
20. Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions. *Nature Machine Intelligence*, 1(5), 206–215.
21. Selbst, A., Boyd, D., Friedler, S., Venkatasubramanian, S., & Vertesi, J. (2019). Fairness and abstraction in sociotechnical systems. *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 59–68.
22. Shapley, L. (1953). A value for n-person games. *Contributions to the Theory of Games*, 2, 307–317.
23. Sunstein, C. (2019). Algorithms, correctives, and the rule of law. *Columbia Law Review*, 119(2), 499–545.
24. Varshney, K. (2019). **Trustworthy machine learning.** IBM Research.
25. Wachter, S., Mittelstadt, B., & Russell, C. (2018). Counterfactual explanations without opening the black box. *Harvard Journal of Law & Technology*, 31(2), 841–887.
26. World Economic Forum. (2020). **Global AI governance report.**
27. Zarsky, T. (2016). The trouble with algorithmic decisions. *Science, Technology & Human Values*, 41(1), 118–132.
28. Goodfellow, I., Bengio, Y., & Courville, A. (2016). **Deep learning.** MIT Press.
29. Russell, S., & Norvig, P. (2021). **Artificial intelligence: A modern approach (4th ed.).** Pearson.
30. Financial Stability Board. (2023). **Artificial intelligence and machine learning in financial services.**