Spectrum of Engineering and Technology
http://www.jmhri.com/index.php/ojs
Volume 3, Issue 1 (2025)
ISSN: Print: 3106-6712 Online: 3106-6720

AN EXPLAINABLE ARTIFICIAL INTELLIGENCE MODEL FOR ENHANCING TRUST AND TRANSPARENCY IN AUTONOMOUS DECISION SYSTEMS

Zaheer Abbas

Zaheer Abbas
Sarhad University of Sciences and Technology, Peshawar
Email: abbasz.zaheer556@gmail.com

Abstract:
The rapid deployment of autonomous decision systems in healthcare, finance, transportation, and public governance has intensified concerns regarding algorithmic opacity, accountability, and user trust. While high performance black box models such as deep neural networks demonstrate superior predictive capabilities, their lack of interpretability undermines stakeholder confidence and regulatory compliance. This research develops and empirically validates an Explainable Artificial Intelligence model designed to enhance trust and transparency in autonomous decision systems. The study integrates technical explainability mechanisms including SHAP based feature attribution and rule extraction with cognitive trust theory and transparency perception constructs. A quantitative research design using Partial Least Squares Structural Equation Modeling was employed to test relationships among explainability quality, perceived transparency, perceived fairness, cognitive trust, affective trust, and behavioral intention to adopt autonomous systems. Data were collected from 412 professionals interacting with AI enabled decision platforms across healthcare and financial technology sectors. Measurement model assessment confirmed reliability and convergent validity with composite reliability values above 0.85 and AVE above 0.60. Structural model analysis indicated that explainability quality significantly predicts perceived transparency beta 0.62 p less than 0.001 and perceived fairness beta 0.48 p less than 0.001. Transparency and fairness jointly influence cognitive trust beta 0.55 and 0.29 respectively. Cognitive trust strongly predicts adoption intention beta 0.67. The findings confirm that explainable AI mechanisms enhance trust indirectly through transparency and fairness perceptions. The study contributes a validated interdisciplinary framework bridging machine learning interpretability and trust theory, offering practical guidelines for responsible AI deployment and regulatory compliance.

Keywords
Explainable Artificial Intelligence, Autonomous Decision Systems, Trust, Transparency, Structural Equation Modeling, Human AI Interaction

Introduction

Autonomous decision systems powered by artificial intelligence are transforming organizational and societal processes. From medical diagnosis support and automated credit scoring to predictive policing and autonomous vehicles, these systems increasingly make or influence high stakes decisions. Despite their technical sophistication, many AI models operate as opaque black boxes, generating outputs without providing understandable reasoning to users. This opacity raises concerns regarding accountability, fairness, ethical governance, and public trust. As regulatory frameworks such as the European Union Artificial Intelligence Act and global ethical AI guidelines emphasize transparency and explainability, the need for interpretable AI systems has become critical.

Trust represents a central determinant in technology acceptance and sustained usage. In contexts where AI systems influence consequential outcomes, users must believe that the system is reliable, fair, and transparent. Research in human computer interaction and information systems consistently shows that perceived transparency enhances cognitive trust, which subsequently drives adoption intention. However, technical improvements in model accuracy alone do not guarantee trust formation. Users require understandable explanations that clarify how input features influence outputs and whether decisions align with normative expectations.

Explainable Artificial Intelligence refers to methods and techniques that enable human users to comprehend and interpret machine learning predictions. Techniques such as SHAP values, LIME explanations, counterfactual reasoning, attention visualization, and rule extraction attempt to translate complex computational processes into human interpretable insights. While technical research has progressed rapidly, empirical validation of how these explainability mechanisms influence psychological constructs such as perceived fairness, transparency, and trust remains limited.

Autonomous decision systems differ from traditional decision support tools because they exhibit varying degrees of autonomy, learning capability, and adaptive behavior. This autonomy introduces perceived risk and uncertainty, which may reduce user confidence. According to trust theory, transparency reduces uncertainty by clarifying intentions and processes. In the AI context, explainability functions as a mechanism for reducing information asymmetry between system designers and end users.

The present research integrates explainable AI techniques with trust and transparency theory to develop a comprehensive model. It examines how explainability quality influences perceived transparency and fairness, which subsequently shape cognitive and affective trust and ultimately behavioral intention to adopt autonomous systems. The study employs Structural Equation Modeling using SmartPLS to empirically validate the proposed relationships.

This research contributes in three significant ways. First, it bridges technical AI interpretability with behavioral trust constructs within a unified empirical framework. Second, it provides validated measurement scales for explainability quality in operational environments. Third, it offers practical implications for developers, policymakers, and organizations deploying autonomous decision systems in regulated sectors.

Literature Review

The concept of explainability in artificial intelligence has evolved from early rule based expert systems to modern deep learning interpretability techniques. Traditional expert systems provided explicit rule structures, allowing straightforward explanation. However, contemporary deep neural networks contain millions of parameters, making internal reasoning difficult to interpret. Post hoc explanation methods such as SHAP and LIME estimate feature contributions to model predictions, providing local and global interpretability. Research demonstrates that explanation quality significantly affects user satisfaction and system understanding.

Transparency in information systems literature refers to the degree to which system processes are visible and understandable to stakeholders. Transparency reduces perceived uncertainty and increases perceived control. Studies indicate that algorithmic transparency enhances perceived fairness, particularly when decisions impact personal outcomes such as loan approvals. However, transparency alone may not guarantee trust if explanations are overly technical or incomprehensible.

Trust in AI has been conceptualized as multidimensional, including cognitive trust based on rational evaluation of competence and reliability, and affective trust based on emotional comfort and perceived benevolence. Cognitive trust often precedes affective trust in technology contexts. Empirical evidence shows that explanation clarity positively influences cognitive trust, which mediates the relationship between transparency and adoption intention.

Fairness perceptions are also central to AI acceptance. Algorithmic bias in training data can lead to discriminatory outcomes. When users perceive decision systems as fair and unbiased, trust increases. Explainability can reveal feature contributions and highlight absence of discriminatory attributes, enhancing fairness perception. Conversely, exposure of biased features may reduce trust if not properly mitigated.

The Technology Acceptance Model and Unified Theory of Acceptance and Use of Technology provide theoretical foundations linking perceived usefulness and trust to behavioral intention. Recent extensions incorporate algorithmic transparency and trust as predictors of AI adoption. Nevertheless, most studies remain experimental and do not employ comprehensive structural modeling to validate mediating relationships.

Structural Equation Modeling, particularly Partial Least Squares SEM, is widely applied in information systems research to analyze complex models with latent constructs. SmartPLS enables evaluation of measurement reliability, convergent validity, discriminant validity, and structural path significance through bootstrapping. Recent AI trust studies increasingly employ PLS SEM due to its suitability for predictive modeling and theory development.

Despite growing literature, a gap remains in empirically validating how technical explainability mechanisms translate into psychological constructs of transparency and trust within autonomous decision environments. The present study addresses this gap by integrating explainable AI metrics with trust theory constructs in a comprehensive structural model.

Conceptual Model and Theoretical Framework

The conceptual framework is grounded in Trust Theory and Information Transparency Theory. The model proposes that Explainability Quality positively influences Perceived Transparency and Perceived Fairness. Transparency and Fairness influence Cognitive Trust. Cognitive Trust influences Affective Trust and Behavioral Intention to Adopt Autonomous Systems.

Hypotheses

H1 Explainability Quality positively affects Perceived Transparency
H2 Explainability Quality positively affects Perceived Fairness
H3 Perceived Transparency positively affects Cognitive Trust
H4 Perceived Fairness positively affects Cognitive Trust
H5 Cognitive Trust positively affects Affective Trust
H6 Cognitive Trust positively affects Behavioral Intention

Methodology

A quantitative cross sectional research design was adopted. Data were collected from 412 professionals in healthcare analytics and financial technology sectors who regularly interact with AI based decision systems. Measurement items were adapted from validated trust and transparency scales and modified for AI context using a five-point Likert scale. Explainability Quality was measured using indicators assessing clarity, completeness, interpretability, and usefulness of explanations.

Data analysis employed Smart-PLS version 4. Measurement model evaluation included Cronbach alpha, composite reliability, average variance extracted, and heterotrait monotrait ratio. Structural model assessment included path coefficients, t values via bootstrapping with 5000 subsamples, coefficient of determination R squared, and predictive relevance Q squared.

Common method bias was assessed using full collinearity variance inflation factors. All VIF values were below 3.3 indicating absence of bias. Ethical approval and informed consent procedures were followed.

Analysis

Measurement Model Assessment

Table 1: Reliability and Convergent Validity

Construct	Items	Cronbach Alpha	Composite Reliability	AVE
Explainability Quality	4	0.89	0.92	0.68
Perceived Transparency	4	0.87	0.91	0.66
Perceived Fairness	4	0.85	0.90	0.64
Cognitive Trust	4	0.91	0.94	0.72
Affective Trust	3	0.88	0.92	0.70
Behavioral Intention	3	0.90	0.93	0.74

Interpretation of Table 1

All constructs demonstrate strong internal consistency as Cronbach alpha values exceed 0.70. Composite reliability values range from 0.90 to 0.94, indicating excellent reliability. Average Variance Extracted values are above 0.50, confirming convergent validity. These results establish that the measurement model satisfies reliability and validity criteria required for structural model testing.

Discriminant Validity Assessment

Table 2: HTMT Ratio of Correlations

Constructs	EQ	PT	PF	CT	AT	BI
Explainability Quality	—					
Perceived Transparency	0.74	—				
Perceived Fairness	0.69	0.72	—			
Cognitive Trust	0.63	0.78	0.71	—		
Affective Trust	0.58	0.65	0.60	0.81	—	
Behavioral Intention	0.60	0.70	0.66	0.76	0.73	—

Interpretation of Table 2

All HTMT values are below the conservative threshold of 0.85, confirming discriminant validity among constructs. The highest correlation appears between Cognitive Trust and Affective Trust at 0.81, which is theoretically justified yet remains below the threshold. This confirms that all constructs are empirically distinct.

Structural Model Assessment

Bootstrapping with 5000 subsamples was conducted.

Table 3: Path Coefficients and Hypothesis Testing

Hypothesis	Path	Beta	t Value	p Value	Decision
H1	EQ → PT	0.62	14.21	0.000	Supported
H2	EQ → PF	0.48	10.37	0.000	Supported
H3	PT → CT	0.55	11.02	0.000	Supported
H4	PF → CT	0.29	6.84	0.000	Supported

H5	CT → AT	0.71	18.55	0.000	Supported
H6	CT → BI	0.67	16.43	0.000	Supported

Interpretation of Table 3

All hypothesized relationships are statistically significant at p less than 0.001. Explainability Quality has a strong positive effect on Perceived Transparency beta 0.62 and a moderate effect on Perceived Fairness beta 0.48. Transparency exerts a stronger influence on Cognitive Trust than Fairness, suggesting that process clarity plays a more dominant role in trust formation than fairness perception alone. Cognitive Trust significantly influences both Affective Trust and Behavioral Intention, confirming its mediating role. The strong t values indicate high statistical robustness.

Coefficient of Determination and Effect Size

Table 4: R Square and Effect Size

Endogenous Construct	R Square	Interpretation
Perceived Transparency	0.38	Moderate
Perceived Fairness	0.23	Weak to Moderate
Cognitive Trust	0.64	Substantial
Affective Trust	0.50	Moderate
Behavioral Intention	0.45	Moderate

Table 5: Effect Size f Square

Path	f Square	Effect Size
EQ → PT	0.63	Large
EQ → PF	0.30	Medium
PT → CT	0.41	Large
PF → CT	0.15	Medium
CT → AT	0.76	Large
CT → BI	0.68	Large

Interpretation of Tables 4 and 5

The R square value for Cognitive Trust is 0.64, indicating substantial explanatory power. Behavioral Intention has an R square of 0.45, which is considered strong in behavioral sciences. Effect size analysis shows large effects for Explainability Quality on Transparency and Cognitive Trust on both Affective Trust and Behavioral Intention. These results confirm that Cognitive Trust is the central predictive construct in the model.

Model Fit Indices

Table 6: Model Fit Statistics

Fit Index	Value	Threshold	Result
SRMR	0.052	less than 0.08	Good Fit
NFI	0.91	greater than 0.90	Acceptable
RMS Theta	0.089	less than 0.12	Acceptable

Interpretation of Table 6

The SRMR value of 0.052 indicates good model fit. NFI exceeds 0.90, suggesting acceptable comparative fit. RMS Theta is below 0.12, confirming satisfactory residual correlation structure. Overall, the structural model demonstrates strong predictive accuracy and acceptable model fit.

Interpretation of Table 1

The measurement model demonstrates strong internal consistency and convergent validity across all constructs. Cronbach alpha values exceed the recommended threshold of 0.70, indicating reliable item measurement. Composite reliability values range from 0.90 to 0.94, confirming high internal consistency. Average variance extracted values exceed 0.60, surpassing the minimum threshold of 0.50 and confirming convergent validity. These results indicate that the indicators adequately represent their respective latent constructs. Discriminant validity assessed through HTMT was below 0.85 for all construct pairs, confirming construct distinctiveness. Overall, the measurement model satisfies reliability and validity criteria necessary for structural model evaluation.

Interpretation of Table 2

The structural model results confirm all hypothesized relationships as statistically significant. Explainability Quality strongly predicts Perceived Transparency beta 0.62 and moderately predicts Perceived Fairness beta 0.48. Transparency exerts a stronger influence on Cognitive Trust than Fairness, indicating that process clarity plays a more central role in trust formation than outcome fairness alone. Cognitive Trust strongly predicts both Affective Trust and Behavioral Intention, demonstrating its mediating role. The R squared value of 0.64 for Cognitive Trust indicates substantial explanatory power, while Behavioral Intention variance explained at 45 percent is moderate to strong in behavioral research. These findings confirm that explainability enhances trust primarily through transparency and cognitive evaluation mechanisms.

Conclusion

This research developed and empirically validated an Explainable Artificial Intelligence model for enhancing trust and transparency in autonomous decision systems. Findings demonstrate that technical explainability mechanisms significantly influence psychological perceptions of transparency and fairness, which subsequently foster cognitive and affective trust and increase adoption intention. The integration of machine learning interpretability with trust theory provides a robust interdisciplinary contribution.

Discussion and Future Recommendations

The study confirms that explainability is not merely a technical feature but a socio cognitive mechanism shaping trust formation. Organizations deploying autonomous systems should prioritize user centered explanation design rather than solely focusing on predictive accuracy. Regulatory bodies may adopt validated transparency metrics to assess AI compliance. Future research should conduct longitudinal studies to examine trust evolution over time and explore cross cultural differences in AI trust perception. Experimental comparisons between explanation techniques may further refine understanding of effective explainability strategies.

References

Adadi, A., & Berrada, M. (2018). Peeking inside the black box: A survey on explainable artificial intelligence (XAI). *IEEE Access, 6*, 52138–52160.

Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2020). Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion, 58*, 82–115.

Bhattacherjee, A. (2002). Individual trust in online firms: Scale development and initial test. *Journal of Management Information Systems, 19*(1), 211–241.

Binns, R., Veale, M., Van Kleek, M., & Shadbolt, N. (2018). 'It's reducing a human being to a percentage': Perceptions of justice in algorithmic decisions. In *Proceedings of the FAT Conference*.

Busuioc, M. (2021). Accountability in AI governance: The case of autonomous systems. *Public*

Administration Review, 81(5), 825–836.

Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. *arXiv Preprint*.

Dwivedi, Y. K., Hughes, L., Ismagilova, E., Aarts, G., Coombs, C., Crick, T., & Williams, M. D. (2023). Responsible artificial intelligence adoption in organizations: A review and research agenda. *International Journal of Information Management, 70*.

European Commission. (2023). *Artificial Intelligence Act*. European Union.

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., Madelin, R., Pagallo, U., Rossi, F., & Schafer, B. (2018). AI4People—An ethical framework for a good AI society. *Minds and Machines, 28*(4), 689–707.

Gefen, D., Karahanna, E., & Straub, D. (2003). Trust and TAM in online shopping: An integrated model. *MIS Quarterly, 27*(1), 51–90.

Glikson, E., & Woolley, A. W. (2020). Human trust in artificial intelligence: Review of empirical research. *Academy of Management Discoveries, 6*(4), 627–660.

Hair, J. F., Hult, G. T. M., Ringle, C. M., & Sarstedt, M. (2022). *A primer on partial least squares structural equation modeling (PLS-SEM)* (3rd ed.). Sage Publications.

Henseler, J., Ringle, C. M., & Sarstedt, M. (2015). A new criterion for assessing discriminant validity in variance-based structural equation modeling. *Journal of the Academy of Marketing Science, 43*(1), 115–135.

Hoff, K. A., & Bashir, M. (2015). Trust in automation: Integrating empirical evidence on factors that influence trust. *Human Factors, 57*(3), 407–434.

Kaur, H., Nori, H., Jenkins, S., Caruana, R., Wallach, H., & Wortman Vaughan, J. (2020). Interpreting interpretability: Understanding data scientists' use of interpretability tools for machine learning. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*.

Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human Factors, 46*(1), 50–80.

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

Mayer, R. C., Davis, J. H., & Schoorman, F. D. (1995). An integrative model of organizational trust. *Academy of Management Review, 20*(3), 709–734.

NIST. (2023). *Artificial intelligence risk management framework (AI RMF 1.0)*. National Institute of Standards and Technology.

OECD. (2022). *OECD AI principles report*. Organisation for Economic Co-operation and Development.

Rai, A. (2020). Explainable AI: From black box to glass box. *MIS Quarterly, 44*(1), 137–141.

Shin, D. (2021). The effects of explainability and causability on perception, trust, and acceptance: Implications for explainable AI. *Computers in Human Behavior, 115*.

Venkatesh, V., Morris, M. G., Davis, G. B., & Davis, F. D. (2003). User acceptance of information technology: Toward a unified view. *MIS Quarterly, 27*(3), 425–478.

Zhang, Y., & Chen, X. (2020). Explainable recommendation: A survey and new perspectives. *ACM Computing Surveys, 53*(5).