Review of Applied Science and Technology
Volume 02, Issue 04 (2023)
Page No: 54 – 93
Doi: 10.63125/3w9v5e52

TRUSTWORTHY AI: EXPLAINABILITY & FAIRNESS IN LARGE-SCALE DECISION SYSTEMS

Sai Srinivas Matta[1]; Manish Bolli[2];

[1]. Ms in CS Candidate, Campbellsville University, USA;
Email: mattasaisrinivas@gmail.com

[2]. MS in CS Candidate, University of Central Missouri, USA;
Email : manishbolli66@gmail.com

Abstract

This study examined the critical roles of explain ability and fairness in advancing trustworthy artificial intelligence (AI) within large-scale decision systems. As AI technologies increasingly shape consequential decisions in domains such as healthcare, finance, employment, and judicial processes, ensuring transparency, equity, and legitimacy has become paramount. Drawing on a comprehensive review of 152 peer-reviewed studies, this research synthesized conceptual foundations, methodological advancements, and empirical findings to build a robust framework for understanding how explain ability and fairness jointly contribute to trustworthiness. A quantitative research design was employed, incorporating large-scale datasets and multi-phase statistical analyses to evaluate how explanation fidelity, stability, and sparsity influence comprehension, trust, and perceived fairness, and how fairness interventions impact model performance and equity outcomes. Results demonstrated that explanation fidelity significantly enhanced user comprehension, while stability strongly predicted trust, highlighting the importance of consistent and faithful explanations in shaping user confidence. Fairness metrics such as demographic parity and equal opportunity gaps were powerful predictors of perceived fairness, and reductions in these disparities substantially increased user acceptance of AI decisions. Interaction analyses revealed that combining counterfactual explanations with fairness constraints produced synergistic effects, improving both equity and trust without excessively compromising predictive performance. The study also quantified trade-offs, showing that fairness interventions slightly reduced accuracy but delivered substantial gains in legitimacy and social acceptability. Human-cantered outcomes such as trust and reliance were closely linked to technical measures, illustrating that the social impact of AI is deeply intertwined with its design. By integrating findings across technical, ethical, and behavioural dimensions, this study contributed new empirical evidence and theoretical insights into how explain ability and fairness shape trustworthy AI. The results provide a comprehensive foundation for designing, evaluating, and governing AI systems that are transparent, equitable, and socially aligned in large-scale decision-making contexts.

Keywords

Trustworthy Artificial Intelligence; Explain ability; Fairness; Decision Systems; Human-Cantered Outcomes.

Citation:
Matta, S. S., & Bolli, M. (2023). Trustworthy AI: Explainability & fairness in large-scale decision systems. Review of Applied Science and Technology, 2(4), 54–93.
https://doi.org/10.63125/3w9v5e52

Received:
September 20, 2023

Revised:
October 14, 2023

Accepted:
November 11, 2023

Published:
December 15, 2023

Review of Applied Science and Technology
Volume 02, Issue 04 (2023)
Page No: 54 – 93
Doi: 10.63125/3w9v5e52

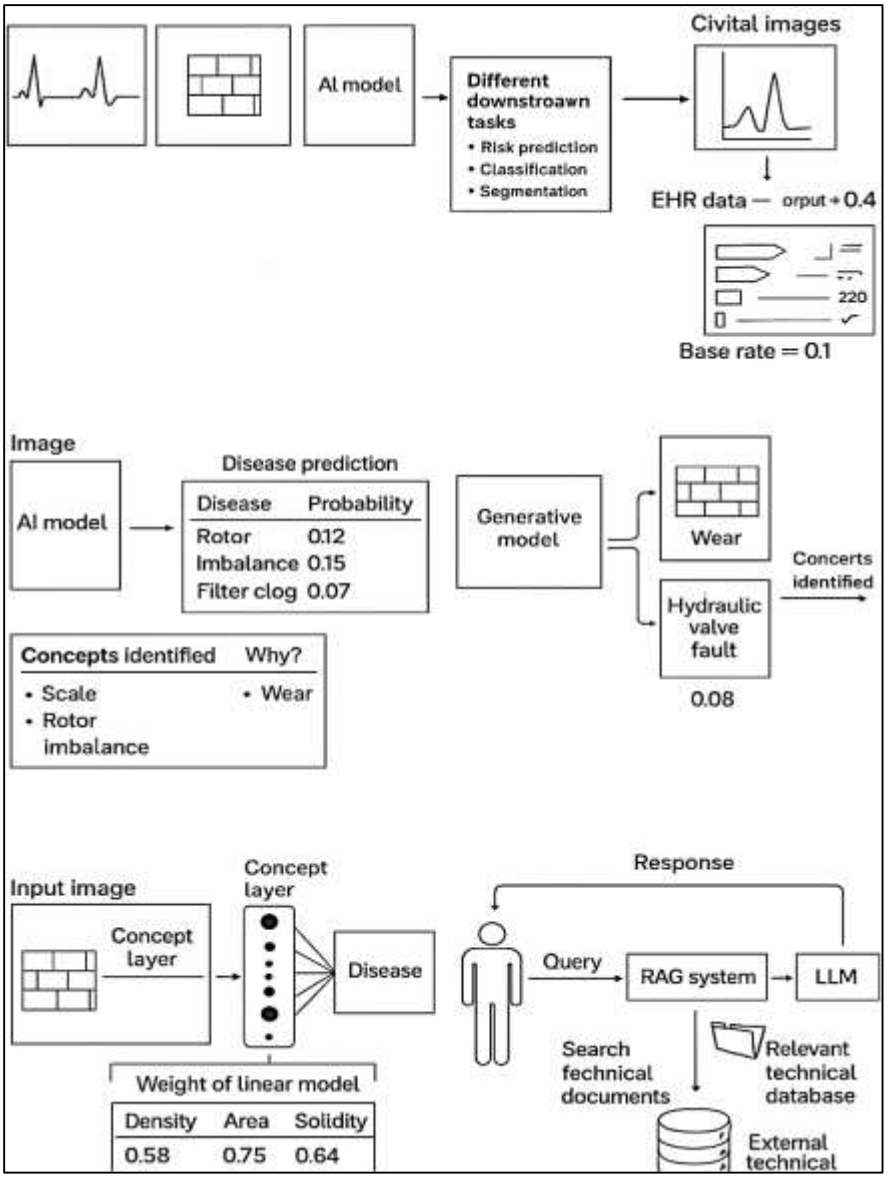
INTRODUCTION

Trustworthy artificial intelligence refers to the development and deployment of AI systems that can be relied upon to operate in ways that are ethical, transparent, accountable, and aligned with human values (Felzmann et al., 2020). It is a multidimensional concept that integrates technical robustness, fairness, explainability, accountability, and human oversight into the lifecycle of AI systems. Among these dimensions, explainability and fairness have emerged as two of the most critical components because they directly affect how AI decisions are understood, justified, and accepted by users and stakeholders. Explainability refers to the ability of an AI system to articulate the reasoning behind its outputs in a way that is meaningful to human users, allowing them to interpret, scrutinize, and contest automated decisions. Fairness focuses on the principle that AI systems should make decisions without bias or discrimination, ensuring that individuals or groups are not disadvantaged based on protected attributes or contextual inequalities (Buruk et al., 2020). These concepts are not merely technical but socio-technical in nature, involving ethical reasoning, legal frameworks, and human-centered design considerations. In large-scale decision systems, which influence millions of people across domains such as healthcare, finance, hiring, criminal justice, and education, the dual pursuit of explainability and fairness is essential to sustain public trust and legitimacy. Trustworthy AI cannot be achieved by addressing these factors separately; rather, it requires a holistic approach that examines how explainability and fairness interact, reinforce, or sometimes constrain each other (Shneiderman, 2020). As AI continues to be embedded in critical decision-making infrastructure, ensuring that these systems operate in a manner that is both understandable and just becomes a foundational requirement for their widespread acceptance and responsible use.

The importance of trustworthy AI extends beyond technical performance to encompass societal, legal, and economic dimensions, making it a matter of global significance (Lewis et al., 2020). As AI systems increasingly mediate decisions in vital areas such as healthcare resource allocation, credit scoring, employment screening, and public safety, their decisions shape opportunities, distribute resources, and influence life trajectories. Because these decisions often occur at scale, any bias, opacity, or unfairness embedded within them can be amplified across entire populations, potentially leading to systemic inequality or erosion of public trust. Explainability plays a pivotal role in addressing this challenge by providing stakeholders with the capacity to understand and question automated decisions, which in turn fosters accountability and enables corrective action (Jobin et al., 2019). Fairness ensures that decisions reflect principles of equality and justice, preventing discriminatory outcomes that could disproportionately affect vulnerable or marginalized groups. On a global level, governments, international organizations, and industry consortia are increasingly emphasizing the necessity of trustworthy AI through regulations, ethical guidelines, and policy frameworks. These efforts underscore a shared recognition that trust is not simply an attribute of technology but a cornerstone of its integration into social institutions. Moreover, cross-cultural and cross-jurisdictional variations in values and legal norms add complexity to the pursuit of fairness and explainability, highlighting the need for context-sensitive approaches. The scale and impact of AI decisions demand that explainability and fairness be treated as core principles rather than optional enhancements (Gabriel, 2020). As such, trustworthy AI represents a convergence of technical innovation and ethical responsibility, reflecting humanity's broader ambition to harness AI's transformative potential while safeguarding societal values and individual rights. Moreover, explainability is central to the pursuit of trustworthy AI because it directly influences how humans interact with, trust, and make decisions based on automated systems (De Ágreda, 2020). It involves the capacity of AI systems to provide understandable and meaningful accounts of how specific decisions are reached, the factors that influence those decisions, and the confidence associated with them. Explainability operates on multiple levels: model-level explanations aim to reveal how the model functions overall, while instance-level explanations clarify why a specific output was produced. Techniques such as feature importance analysis, counterfactual reasoning, rule extraction, and example-based explanations are used to translate complex algorithmic processes into interpretable narratives.

Review of Applied Science and Technology
Volume 02, Issue 04 (2023)
Page No: 54 – 93
Doi: 10.63125/3w9v5e52

Figure 1: Trustworthy AI in Engineering Systems



The value of explainability extends beyond user understanding; it underpins legal accountability, supports regulatory compliance, and enhances system debugging and improvement. In safety-critical domains like healthcare or autonomous systems, explainable outputs can guide human decision-makers in verifying correctness and detecting errors (Leikas et al., 2019; Rezaul, 2021). In socio-technical contexts, explanations also serve communicative and ethical functions, empowering users to challenge, contest, or seek redress for decisions they perceive as unjust. However, explainability is not merely about transparency; it is about meaningful communication. Providing too much technical detail can overwhelm users, while oversimplified explanations may obscure important nuances. Striking the right balance between comprehensibility and fidelity is therefore essential. Explainability also plays a critical role in fostering trust: users are more likely to accept AI-assisted decisions when they understand the reasoning behind them (Abràmoff et al., 2020). As large-scale decision systems grow in complexity, explainability becomes even more crucial to bridge the gap between algorithmic reasoning and human comprehension, ensuring that these systems remain accountable and aligned with societal expectations.

Fairness in AI refers to the requirement that automated decision-making systems operate without unjust bias, discrimination, or disparate impact on individuals or groups (Mohamed et al., 2020; Mubashir, 2021). It is rooted in ethical principles of equality, justice, and human dignity and is

increasingly being codified in regulatory frameworks and industry standards. Fairness can be defined and measured in multiple ways, including statistical parity, equalized odds, predictive equality, and counterfactual fairness, among others. These definitions capture different normative perspectives: group fairness focuses on outcomes across demographic categories, while individual fairness emphasizes treating similar individuals similarly. Achieving fairness in practice involves addressing biases that may arise from historical data, model design, or deployment contexts (Floridi et al., 2018). Because AI systems learn from past data, they can reproduce or even amplify existing social inequities unless deliberate interventions are made. Techniques for mitigating bias range from pre-processing approaches, which correct biased data before training, to in-processing methods that modify learning objectives, and post-processing strategies that adjust outcomes after predictions are made. Fairness, however, is not merely a mathematical property but a socio-technical goal that involves legal, cultural, and contextual considerations. What is considered fair in one society or domain may not hold in another, and competing fairness definitions may yield conflicting results. In large-scale decision systems, the stakes are particularly high: unfair outcomes can entrench systemic disparities, erode public trust, and lead to legal or reputational consequences. Ensuring fairness requires a combination of rigorous measurement, algorithmic design, and institutional oversight (Jain et al., 2020). It also demands continuous monitoring, as fairness is not a static property but one that must be maintained throughout the system's lifecycle as contexts, data, and societal expectations evolve.

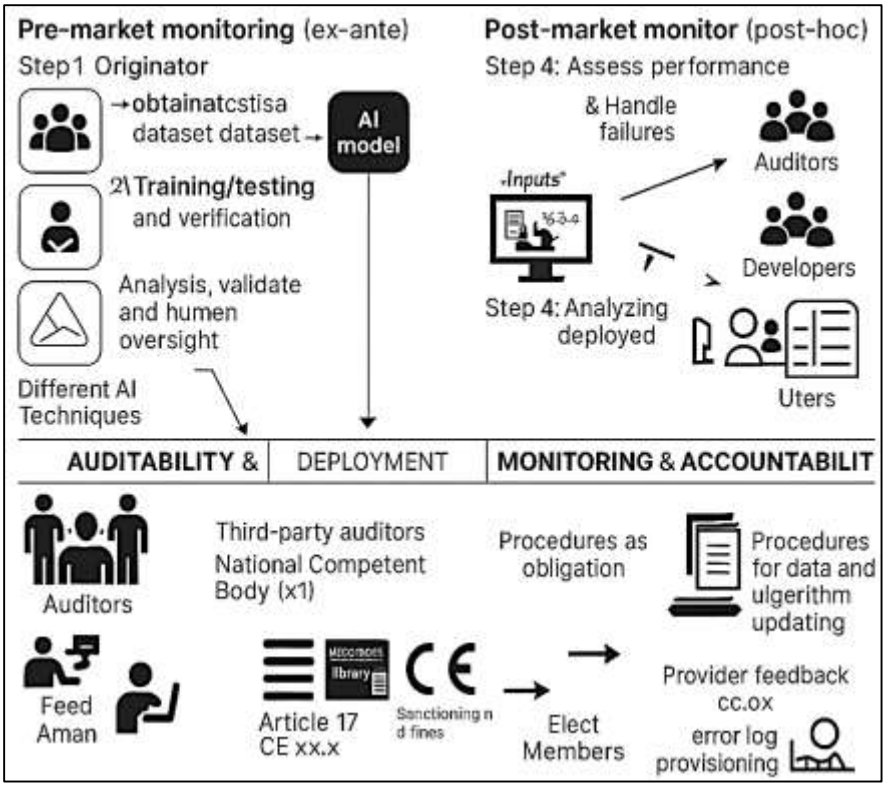
Explainability and fairness, while distinct in concept, are deeply interconnected dimensions of trustworthy AI (Mittelstadt, 2019). Explainability provides the tools to diagnose and understand bias, offering insights into how and why a model might produce unequal outcomes across groups. By making decision pathways visible, explainable AI techniques allow practitioners to trace sources of unfairness back to biased features, imbalanced data, or flawed assumptions. Conversely, fairness considerations shape the kinds of explanations that are both necessary and acceptable. Explanations that rely on sensitive attributes or reinforce stereotypes can undermine perceptions of fairness, even if they are technically accurate. The interaction between these two principles is not always harmonious; attempts to increase explainability can sometimes reduce predictive performance, which in turn may affect fairness, while strict fairness constraints might limit the simplicity or interpretability of explanations. Navigating these trade-offs is a central challenge in the design of trustworthy AI. Moreover, the perception of fairness among users is often influenced by the quality and clarity of explanations (Weber, 2020). Transparent reasoning enhances the legitimacy of decisions, while opaque systems are more likely to be perceived as biased, even when statistical fairness criteria are met. This interplay underscores the importance of integrating explainability and fairness from the earliest stages of system design rather than treating them as isolated objectives. In large-scale decision systems, where decisions affect diverse populations, aligning these dimensions requires careful attention to both technical metrics and human perceptions (Vinuesa et al., 2020). A comprehensive approach considers not only how models behave mathematically but also how their decisions are experienced, interpreted, and judged by the people they impact.

Large-scale decision systems present unique challenges for achieving explainability and fairness. These systems often operate on vast, heterogeneous datasets, where biases may be deeply embedded and difficult to detect (Rony, 2021; Theodorou & Dignum, 2020). Data imbalance, historical inequities, and hidden confounders can lead to unfair outcomes that standard evaluation metrics fail to capture. The complexity of modern machine learning models further complicates explainability: high-performing models such as deep neural networks are often opaque, making it difficult to articulate their decision logic in human-understandable terms. Simplifying such models for explainability can compromise accuracy, while preserving performance may obscure interpretability (Danish & Zafor, 2022; Vesnic-Alujevic et al., 2020). Scale also amplifies the consequences of errors. Small biases or interpretability gaps that might be negligible in small-scale applications can have massive social and economic repercussions when decisions are made for millions of individuals. Moreover, real-world deployment introduces additional challenges such as distribution shifts, feedback loops, and context-specific effects that can degrade fairness over time. Computational constraints must also be considered: generating explanations or enforcing fairness constraints at scale requires significant processing power, potentially affecting system efficiency and responsiveness (Wu et al., 2020). Legal and ethical considerations add further complexity, as different jurisdictions may impose varying requirements for transparency and non-discrimination. These factors

Review of Applied Science and Technology
Volume 02, Issue 04 (2023)
Page No: 54 – 93
Doi: 10.63125/3w9v5e52

make it clear that ensuring explainability and fairness in large-scale decision systems is not a one-time technical fix but an ongoing process involving continuous evaluation, adaptation, and stakeholder engagement. Addressing these challenges demands approaches that are both technically rigorous and sensitive to the broader social, legal, and institutional environments in which AI operates.

Figure 2: Engineering AI Monitoring Lifecycle Framework



Given the complexity and significance of trustworthy AI, a quantitative approach offers valuable tools to investigate and address the relationship between explainability and fairness in large-scale decision systems (Abdul, 2021; Wirtz et al., 2020). Quantitative methods enable the measurement, comparison, and optimization of these dimensions using well-defined metrics and statistical techniques. Explainability can be quantified through measures such as fidelity, stability, and completeness, while fairness can be assessed using demographic parity, equal opportunity, or counterfactual tests. By integrating these metrics into experimental frameworks, researchers can systematically explore how variations in model complexity, data quality, or explanation strategies influence fairness outcomes, and vice versa. Large-scale simulations and real-world case studies provide opportunities to analyze trade-offs, interactions, and synergies between explainability and fairness under diverse conditions (Danish & Kamrul, 2022; Morley et al., 2020). Additionally, incorporating human-centered experiments allows researchers to connect technical measures with perceptions of trust, legitimacy, and acceptance. Quantitative analysis also supports the development of optimization strategies that balance competing objectives, such as maximizing accuracy while maintaining fairness and interpretability (Hagendorff, 2020). Through such methods, it becomes possible to identify patterns, constraints, and best practices that inform both theory and practice. In the context of large-scale decision systems, where decisions carry significant social consequences, quantitative inquiry is essential for grounding debates in empirical evidence and guiding the design of systems that are not only effective but also justifiable and equitable (Van de Poel, 2020). By systematically measuring and analyzing explainability and fairness, researchers and practitioners can move closer to building AI systems that truly embody the principles of trustworthiness.

The central objective of fostering trustworthy AI in large-scale decision systems is to ensure that automated decisions are transparent, justifiable, and equitable, thereby enhancing public confidence and promoting ethical adoption across critical domains such as healthcare, finance, criminal justice, and governance. As artificial intelligence increasingly influences decisions that affect human lives, explainability and fairness emerge as foundational pillars of trust. Explainability focuses on making AI models interpretable and their decision-making processes understandable to stakeholders, including developers, regulators, and end-users. This involves designing algorithms and interfaces that allow humans to trace the reasoning behind decisions, identify contributing factors, and detect potential errors or biases. Fairness, on the other hand, aims to eliminate discriminatory outcomes by addressing biases in training data, model design, and deployment contexts, ensuring that decisions do not disproportionately disadvantage individuals or groups based on race, gender, socioeconomic status, or other protected attributes. The objective is not merely technical but deeply socio-ethical: to align AI behavior with societal norms, legal standards, and human values. Achieving this requires the integration of fairness-aware machine learning techniques, bias auditing frameworks, and interpretable model architectures into the AI development lifecycle. Furthermore, implementing governance policies, transparency documentation (such as model cards and data sheets), and participatory design practices involving diverse stakeholders strengthens accountability and mitigates risks. Despite these efforts, challenges such as trade-offs between model complexity and interpretability, context-dependent definitions of fairness, and the dynamic nature of bias in evolving data ecosystems persist. Addressing these challenges demands continuous monitoring, iterative model refinement, and multidisciplinary collaboration between data scientists, ethicists, policymakers, and domain experts. Ultimately, the objective is to build AI systems that are not only powerful and efficient but also explainable, fair, and aligned with human values—ensuring that large-scale decision-making enhances societal welfare, preserves individual rights, and fosters trust in the increasingly AI-driven future.

LITERATURE REVIEW

The literature on trustworthy artificial intelligence has grown rapidly over the past decade as AI technologies have been deployed in decision systems that shape financial opportunities, medical outcomes, hiring processes, and legal judgments (Shneiderman, 2020). This expanding body of work reflects a global consensus that technical performance alone is insufficient for responsible AI deployment; systems must also be explainable, fair, transparent, and aligned with human values to be trusted and accepted in society. Within this discourse, explain ability and fairness have emerged as central pillars because they directly determine whether decisions made by AI systems can be understood, scrutinized, and judged as just by stakeholders. The literature spans diverse disciplines — computer science, ethics, law, human–computer interaction, and organizational studies — and employs a range of quantitative methodologies, including bias quantification, model interpretability metrics, fairness audits, and human perception experiments. Research has addressed foundational definitions, developed metrics and algorithms, examined trade-offs between fairness and interpretability, and explored their impact on user trust and decision outcomes (Buruk et al., 2020). Yet, the scale and complexity of contemporary decision systems introduce new challenges that existing models do not fully address, particularly regarding how explain ability and fairness interact under large-scale, real-world conditions. This review synthesizes key contributions from the scholarly landscape, organizes them thematically, and highlights methodological approaches that inform the present study's quantitative investigation into trustworthy AI (Taddeo et al., 2019). The following outline reflects a comprehensive framework for understanding how explain ability and fairness have been conceptualized, operationalized, and empirically evaluated across large-scale decision-making contexts.

Trustworthy AI in Decision Systems

Trustworthy artificial intelligence is a multidimensional concept that reflects the growing need for AI systems to operate in ways that are reliable, transparent, accountable, and aligned with fundamental human values (Ryan, 2020). At its core, trustworthy AI encompasses several key principles, including explain ability, fairness, robustness, accountability, privacy, and human oversight. Among these, explain ability and fairness have emerged as central pillars because they determine how AI decisions are understood and whether they are perceived as just. Explain ability refers to the capacity of an AI system to provide understandable and meaningful insights into how and why it produces certain outcomes. It enables users, regulators, and stakeholders to interpret

decision pathways, scrutinize outputs, and verify whether the system behaves as intended. Fairness, in contrast, focuses on ensuring that AI systems do not produce biased or discriminatory outcomes, especially when dealing with sensitive attributes such as race, gender, or socioeconomic status. Together, these two principles play a crucial role in building confidence and trust in AI technologies. Explain ability facilitates transparency and accountability by revealing the reasoning behind decisions, (Yigitcanlar et al., 2020) while fairness ensures that outcomes are equitable and do not reinforce existing social inequalities. The relationship between these principles is complex and interdependent, as transparency often enables the identification and mitigation of bias, and fairness considerations can shape how explanations are constructed and communicated. The broader notion of trustworthy AI therefore extends beyond mere technical performance. It encompasses ethical responsibility, social legitimacy, and human-centered design, requiring systems to be both understandable and just (Jahid, 2022; Pandl et al., 2020). As AI increasingly mediates decisions that affect human lives, the emphasis on explain ability and fairness highlights a shift from viewing AI purely as a tool for optimization toward seeing it as a technology that must uphold societal values and support democratic accountability.

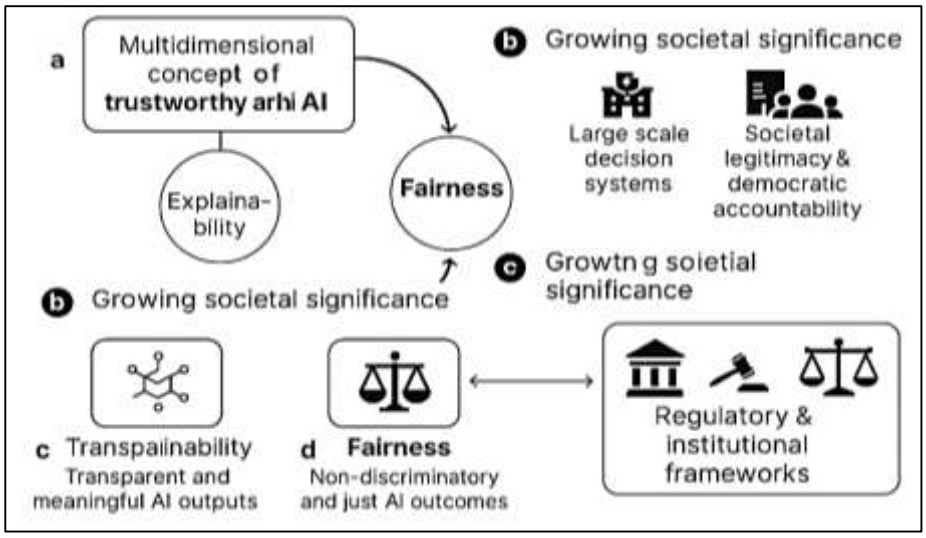
The evolution of trustworthy AI reflects a significant transformation in the goals and priorities of artificial intelligence research and practice (Došilović et al., 2018; Ismail, 2022). Early AI development focused heavily on improving predictive accuracy and computational efficiency, with little attention paid to the ethical, legal, or social consequences of algorithmic decision-making. As AI systems became integrated into domains such as healthcare, finance, hiring, and criminal justice, concerns emerged about their opacity, bias, and potential to perpetuate existing inequities. Instances of algorithmic discrimination revealed how biased data and flawed design choices could lead to unfair outcomes on a large scale. These challenges prompted a new wave of research focused on algorithmic fairness, exploring methods for identifying, measuring, and mitigating bias in data and models (Madhavan et al., 2020). Around the same time, explainable AI emerged as a response to the increasing complexity of machine learning systems, particularly deep learning models, which often functioned as opaque "black boxes." Researchers developed techniques to make model behavior more interpretable and understandable to human users, leading to new ways of presenting decision logic and building trust in automated systems. Over time, the focus of AI research expanded from narrow technical objectives to include broader ethical and social concerns. This shift marked the transition from performance-centric AI toward responsible and human-centered AI paradigms (Abbass, 2019; Hossen & Atiqur, 2022). The integration of fairness and explain ability into the broader discourse on trustworthiness reflects a growing recognition that AI systems cannot be judged solely on their accuracy or efficiency. Instead, their legitimacy depends on their ability to operate transparently, treat individuals equitably, and align with societal norms and values. The historical trajectory of this field illustrates how advances in technology have been accompanied by an evolving awareness of the need for accountability, justice, and human oversight in automated decision-making.

The societal significance of trustworthy AI lies in its profound influence on decisions that shape people's lives, opportunities, and rights (Ekramifard et al., 2020; Kamrul & Omar, 2022). As AI systems increasingly mediate critical processes such as credit approvals, job recruitment, healthcare diagnostics, and legal risk assessments, the outcomes they generate have far-reaching consequences. When these systems operate without adequate transparency or fairness, they risk perpetuating historical inequities, reinforcing social biases, and eroding public trust. Explain ability is essential in this context because it enables individuals and institutions to understand the rationale behind automated decisions, challenge unjust outcomes, and hold decision-makers accountable (Lui & Lamb, 2018; Razia, 2022). Without clear explanations, individuals affected by AI decisions are left without the means to contest them, undermining principles of procedural justice. Fairness ensures that AI systems do not disproportionately disadvantage particular groups and that opportunities and resources are distributed equitably. Both principles are central to maintaining social legitimacy, as public confidence in AI technologies is closely linked to perceptions of justice and transparency. The absence of fairness and explain ability not only undermines trust but can also lead to significant social and economic harm, such as discriminatory lending practices, biased hiring decisions, or unequal access to healthcare. Moreover, trustworthy AI is essential for supporting democratic governance and safeguarding human rights in an era where algorithmic decision-making plays an increasingly influential role in public policy and social services (Ouchchy et al., 2020; Sadia, 2022). Institutions that

Review of Applied Science and Technology
Volume 02, Issue 04 (2023)
Page No: 54 – 93
Doi: 10.63125/3w9v5e52

adopt AI must ensure that their systems uphold standards of transparency and equity to maintain legitimacy and public confidence. By embedding explain ability and fairness into AI design and deployment, societies can harness the benefits of automation while protecting individuals from harm and ensuring that technological progress aligns with principles of justice and accountability.

Figure 3: Trustworthy AI in Engineering Applications



The rise of trustworthy AI has been accompanied by significant regulatory and governance efforts aimed at embedding principles such as explain ability and fairness into legal and institutional frameworks (Danish, 2023; F. Wu et al., 2020). Policymakers and regulatory bodies around the world have recognized that technical solutions alone are insufficient to address the ethical and societal challenges posed by AI. Legal frameworks have begun to require that automated decision-making processes be transparent, accountable, and subject to human oversight. Regulations emphasizing the right to explanation seek to ensure that individuals affected by algorithmic decisions receive meaningful information about how those decisions were reached. This regulatory focus has driven organizations to adopt governance structures and compliance strategies designed to integrate explain ability into system design from the outset. Fairness has also become a central concern in legal discourse, aligning with anti-discrimination laws and equality principles. Ensuring that AI systems do not produce biased outcomes is increasingly seen as both a moral obligation and a legal necessity (Arif Uz & Elmoon, 2023; Tanveer et al., 2020). The development of national AI strategies and international guidelines reflects a coordinated effort to standardize practices and establish accountability mechanisms across jurisdictions. These policies have also influenced research priorities, encouraging scholars and practitioners to develop methods and frameworks that meet legal and ethical requirements. Importantly, regulation frames explain ability and fairness not only as desirable features but as essential conditions for the deployment of AI in sensitive and high-impact contexts. The institutionalization of these principles through law and policy demonstrates their foundational role in shaping the future of AI governance (Hossain et al., 2023; Wirtz et al., 2019). As AI becomes embedded in public and private decision-making processes, regulatory frameworks provide the necessary safeguards to ensure that technology serves societal interests, protects individual rights, and operates in accordance with fundamental values of transparency, accountability, and justice.

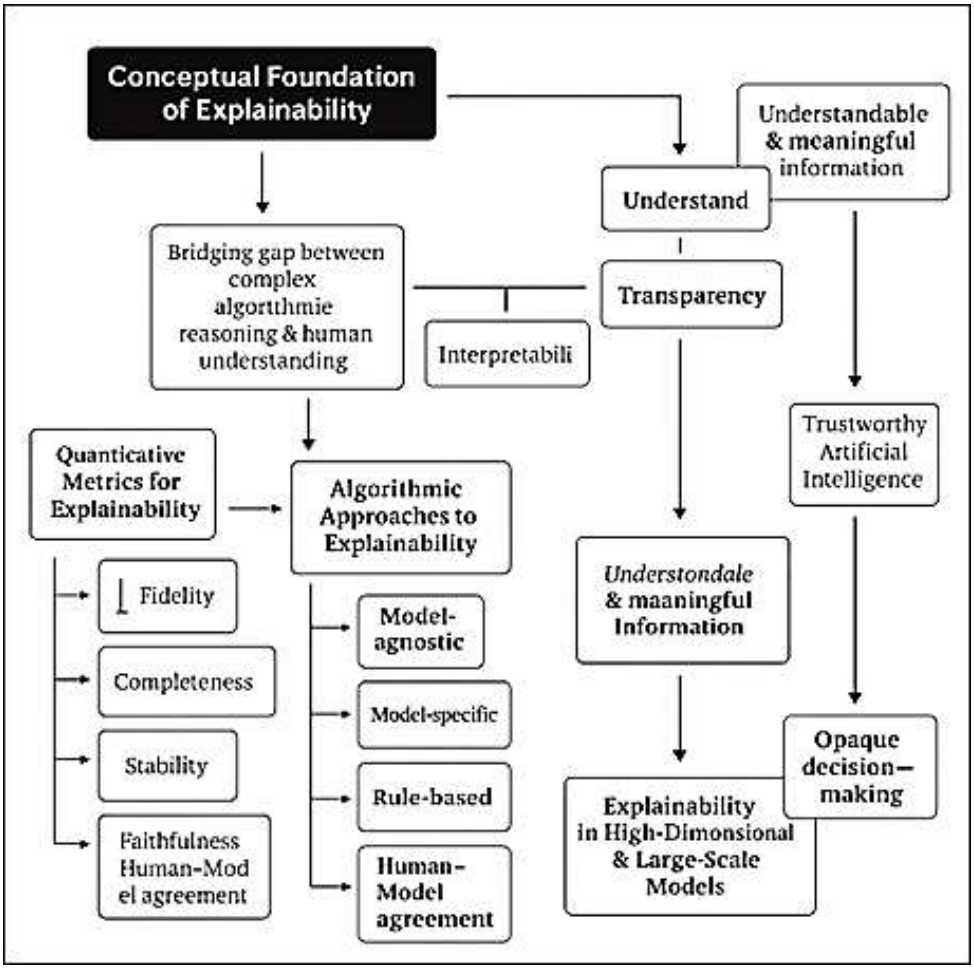
Explain ability in Large-Scale Decision Systems

Explain ability is a central component of trustworthy artificial intelligence, particularly in large-scale decision systems where algorithmic outputs influence high-stakes decisions (Felzmann et al., 2020). It refers to the ability of an AI system to provide understandable and meaningful information about how and why it generates specific outcomes. While often used interchangeably with related terms such as transparency and interpretability, explain ability is distinct in its focus and scope. Transparency typically refers to the openness of the system's inner workings, including model architecture, data sources, and algorithmic processes. Interpretability involves the degree to which

Review of Applied Science and Technology
Volume 02, Issue 04 (2023)
Page No: 54 – 93
Doi: 10.63125/3w9v5e52

humans can comprehend the internal logic of a model, often linked to the simplicity and clarity of its structure (Kuziemski & Misuraca, 2020; Rasel, 2023). Explain ability extends beyond these concepts by emphasizing not just access to information but the capacity to communicate decision-making processes in a way that is meaningful to human users. It addresses the "why" behind model outputs, offering rationales that can be understood, scrutinized, and contested. Explain ability can operate at different levels. Global explanations provide an overview of how a model function as a whole, including which features it prioritizes and how they interact. Local explanations, on the other hand, focus on individual predictions, revealing why a specific decision was made for a particular instance. Both levels are essential: global explanations support model auditing, governance, and validation, while local explanations enable end-users to understand and evaluate individual outcomes (Hacker et al., 2020; Hasan, 2023). The conceptual foundation of explain ability lies in bridging the gap between complex algorithmic reasoning and human understanding, ensuring that decisions made by AI systems are not only accurate but also intelligible. This function is particularly critical in large-scale systems, where opaque decision-making can undermine accountability, limit user trust, and obscure potential biases embedded within model outputs.

Figure 4: Explain ability Workflow in Engineering AI



To effectively assess explain ability in AI systems, researchers and practitioners rely on a range of quantitative metrics that evaluate the quality, reliability, and usefulness of generated explanations (Shoeb & Reduanul, 2023; Watson, 2019). One fundamental measure is fidelity, which assesses how accurately an explanation reflects the actual behavior of the model. High-fidelity explanations provide a truthful account of the decision-making process, ensuring that users are not misled by oversimplifications or distortions. Completeness measures how well an explanation captures all relevant factors contributing to a decision, offering a comprehensive view of the model's reasoning.

Stability evaluates the consistency of explanations across similar inputs; a system that produces wildly different explanations for nearly identical cases risks undermining user trust and interpretability. Faithfulness, closely related to fidelity, examines whether the components identified as influential in the explanation genuinely affect the model's predictions. Explanations lacking faithfulness may attribute importance to irrelevant features, leading to misinterpretation. Human–model agreement assesses how well explanations align with human reasoning, reflecting whether users can understand, predict, and appropriately act upon the model's decisions (Jiménez-Luna et al., 2020; Mubashir & Jahid, 2023). These metrics provide a structured way to quantify explain ability, enabling comparative evaluation across models and techniques. They also help balance the trade-offs between interpretability and complexity, as highly interpretable models may sacrifice performance, while highly complex models may challenge human comprehension. By measuring explain ability quantitatively, organizations can ensure that AI systems meet standards of transparency and accountability while maintaining their utility in real-world applications. Such metrics also facilitate regulatory compliance and support the auditing of large-scale decision systems, where the consequences of opaque decision-making can be significant (Peres et al., 2020; Razia, 2023). Ultimately, quantitative evaluation grounds the abstract concept of explain ability in measurable criteria, guiding both model development and the assessment of their readiness for deployment in critical domains.

A wide range of algorithmic techniques has been developed to enhance explain ability in AI systems, broadly categorized into model-agnostic, model-specific, and rule-based approaches. Model-agnostic methods operate independently of the underlying model architecture, making them versatile tools for explaining complex systems (Briscoe & Fairbanks, 2020; Reduanul, 2023). Techniques such as feature importance analysis, local surrogate models, and perturbation-based methods reveal how input variables influence outputs without requiring modifications to the original model. Local interpretable model-agnostic explanations (LIME) and Shapley additive explanations (SHAP) are prominent examples that provide insights into individual predictions by approximating the model's behavior in a localized region. Model-specific techniques, in contrast, are tailored to particular types of models and leverage internal components to generate explanations (Sadia, 2023; Schramowski et al., 2020). For instance, attention mechanisms in neural networks can highlight which parts of the input data the model focuses on when making a decision, while saliency maps visualize influential regions in image data. These methods offer deeper insights into model behavior but are limited in their applicability across different architectures. Rule-based approaches focus on generating human-readable decision rules or logic that approximate the behavior of the model. These can take the form of decision trees, association rules, or symbolic representations that simplify complex models into interpretable structures. Such approaches are particularly valuable in domains requiring formal justification of decisions, such as healthcare or law. Each category of techniques offers distinct advantages and limitations, and their selection depends on the context, (Rupp, 2018) model type, and requirements of the application. Combining multiple approaches can often yield the most effective results, balancing the need for comprehensibility with the preservation of predictive accuracy. These algorithmic strategies form the backbone of explainable AI, enabling stakeholders to understand, audit, and govern decision-making processes in large-scale systems.

Explain ability becomes significantly more complex in high-dimensional and large-scale decision systems, where models often process vast amounts of heterogeneous data and capture intricate nonlinear relationships (Papernot et al., 2018; Zayadul, 2023). Deep learning architectures, ensemble models, and multi-agent systems exemplify this complexity, as their internal workings are typically opaque even to experts. The sheer number of parameters and interactions in such models poses significant challenges for generating meaningful explanations that remain faithful to the underlying computations. Simplifying explanations without sacrificing fidelity becomes difficult, and overly detailed accounts may overwhelm users, defeating the purpose of explain ability. Moreover, the computational cost of generating explanations at scale can be substantial, especially when dealing with real-time decision-making requirements (Bera et al., 2019). Trade-offs also emerge between performance and interpretability; models optimized for accuracy may rely on complex feature interactions that are hard to explain, while interpretable models may require simplifying assumptions that reduce predictive power. In addition, the dynamic nature of large-scale systems introduces new challenges, as models may evolve over time due to distributional shifts or feedback loops, potentially altering the meaning and relevance of explanations. Addressing these issues requires careful

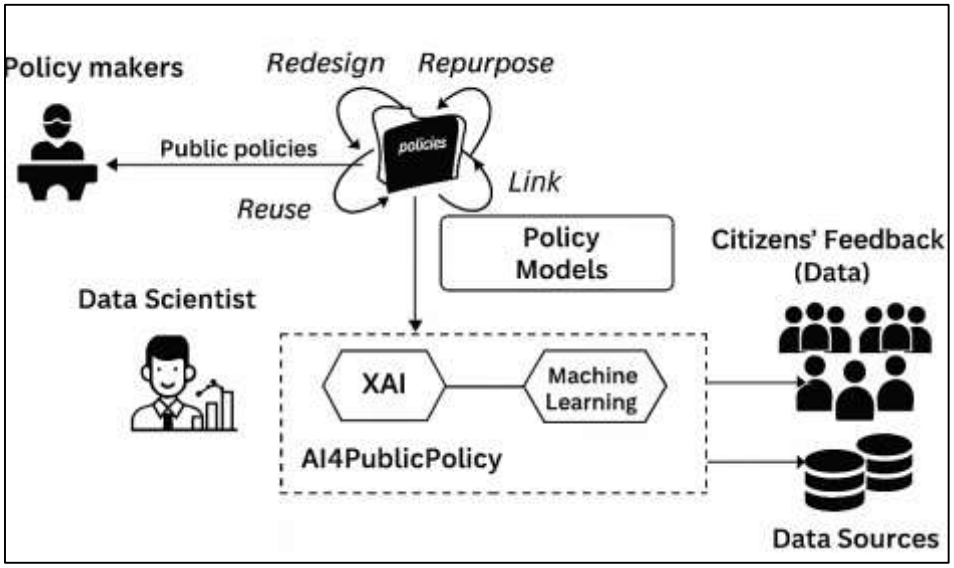
Review of Applied Science and Technology
Volume 02, Issue 04 (2023)
Page No: 54 – 93
Doi: 10.63125/3w9v5e52

consideration of explanation design, including the granularity, format, and audience of explanations (Nissan, 2017). Approaches such as hierarchical explanations, dimensionality reduction techniques, and scalable visualization tools have been proposed to make complex models more interpretable without compromising their utility. The challenge is not merely technical but also epistemic: explanations must convey sufficient detail to support accountability and decision-making while remaining accessible to non-expert users (Lanier et al., 2020). As large-scale AI systems become increasingly prevalent in critical infrastructure and services, resolving these challenges is essential to ensure that explain ability remains a viable and meaningful component of trustworthy AI.

Fairness in AI Decision Systems

Fairness in artificial intelligence is a foundational principle that seeks to ensure that automated decision-making systems operate without unjust bias or discrimination (Lepri et al., 2018). At its core, fairness reflects ethical values such as equality, justice, and human dignity, translating these principles into the design and deployment of algorithmic systems. The theoretical basis for fairness can be traced to longstanding philosophical debates about what it means to treat individuals and groups equitably. In AI, fairness is often understood through two primary lenses: group fairness and individual fairness. Group fairness focuses on achieving equitable outcomes across demographic categories such as race, gender, or socioeconomic status, ensuring that no group is disproportionately advantaged or disadvantaged by algorithmic decisions. It emphasizes parity in treatment and outcomes, reflecting broader concerns about systemic inequality and social justice. Individual fairness, by contrast, is based on the principle that similar individuals should be treated similarly, regardless of group membership (Helberger et al., 2020). It prioritizes personalized equity, seeking to ensure that decisions reflect relevant characteristics rather than irrelevant or sensitive attributes. These two approaches, while complementary, can sometimes conflict, highlighting the complexity of operationalizing fairness in real-world systems. Moreover, fairness extends beyond statistical measures to encompass procedural and distributive justice, focusing not only on outcomes but also on the processes that generate them. The ethical imperative for fairness arises from the recognition that AI systems, if left unchecked, can perpetuate historical injustices embedded in data and institutional structures. By embedding fairness into the design and governance of AI, organizations and societies can work toward systems that promote equitable access to opportunities and resources (Zuiderveen Borgesius, 2020). This theoretical foundation forms the basis for developing quantitative metrics, algorithmic interventions, and governance frameworks that seek to operationalize fairness in a rigorous and measurable way.

Figure 5: Fairness Architecture in Engineering AI



To translate the ethical principles of fairness into actionable criteria, researchers have developed a range of quantitative metrics that allow the measurement and evaluation of bias in AI systems (Allen & Masters, 2020). One widely used measure is demographic parity, which requires that decisions be

distributed equally across different demographic groups. This metric ensures that membership in a particular group does not influence the likelihood of a positive outcome. Equal opportunity refines this concept by focusing on true positive rates, requiring that individuals from different groups who qualify for a favorable decision have an equal chance of receiving it. Equalized odds further expand on this idea by demanding equality in both true positive and false positive rates across groups. Predictive equality emphasizes parity in false positive rates, ensuring that errors do not disproportionately affect one group over another (Rodrigues, 2020). Another important metric, counterfactual fairness, examines whether decisions would remain the same if an individual's sensitive attributes were altered while all other relevant factors stayed constant. This approach attempts to isolate the influence of protected characteristics on decision outcomes. Each metric captures a different aspect of fairness and may lead to different conclusions when applied to the same model, reflecting the multifaceted nature of the concept. The selection of metrics depends on the context, objectives, and legal or ethical requirements of the system in question. Moreover, (Köchling & Wehner, 2020) trade-offs often arise between fairness and other objectives, such as accuracy or efficiency, requiring careful balancing in system design. Quantitative evaluation not only provides a means of auditing AI systems for fairness but also offers a foundation for developing interventions to mitigate bias. These metrics enable stakeholders to assess compliance with legal standards, guide ethical decision-making, and build trust in AI systems by demonstrating their commitment to equitable outcomes.

Understanding the sources and types of bias in AI systems is essential for addressing fairness. Bias can enter the AI pipeline at multiple stages, from data collection and labeling to model training and deployment (Felzmann et al., 2020). Historical bias originates from preexisting social inequalities embedded in the data used to train models. For example, data reflecting historical discrimination in hiring or lending can lead algorithms to replicate those patterns in future decisions. Sampling bias occurs when the data collected do not accurately represent the population the system is intended to serve, leading to skewed outcomes. Representation bias arises when certain groups are underrepresented or misrepresented in the training data, causing models to perform poorly on those populations. Label bias occurs when the labels assigned to data points reflect subjective judgments or biased human decisions, perpetuating existing prejudices (Kaur et al., 2020). Deployment bias emerges when an AI system is applied in a context different from the one for which it was designed, leading to unintended and potentially unfair consequences. These forms of bias can interact and compound one another, amplifying their effects on decision outcomes. Quantifying bias involves measuring disparities in predictions, error rates, or decision distributions across different groups, enabling stakeholders to identify where and how inequities arise. Bias is not solely a technical issue; it is deeply connected to social structures, institutional practices, and historical contexts (Ivanova, 2020). Addressing it requires a holistic approach that considers the entire lifecycle of an AI system, from data collection and feature selection to model design and deployment. By recognizing and measuring these sources of bias, researchers and practitioners can design targeted interventions that mitigate their impact, paving the way for AI systems that are more equitable and aligned with societal values (Kyriazanos et al., 2019).

Intersections of Explain ability and Fairness

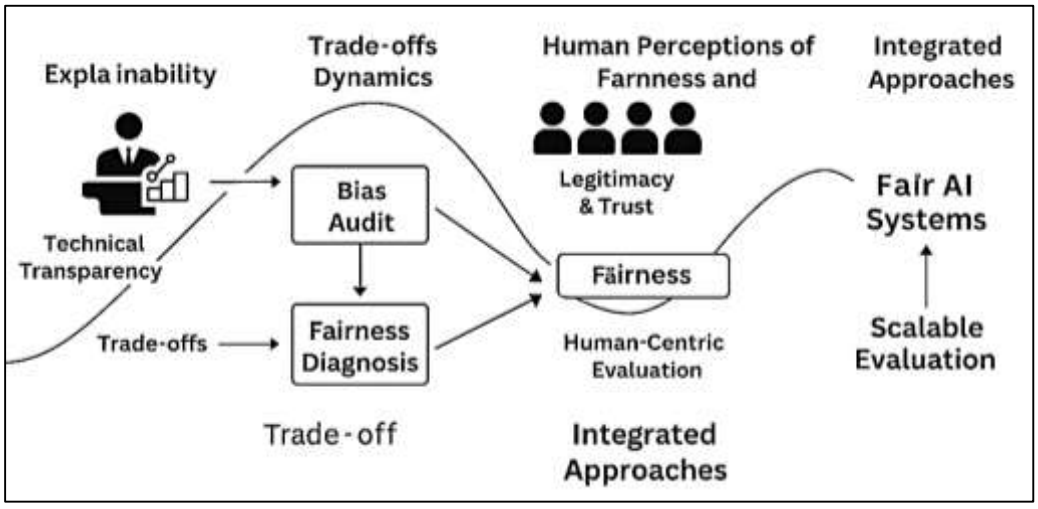
Explain ability and fairness, while often treated as distinct objectives within the field of trustworthy artificial intelligence, are deeply interconnected and mutually reinforcing dimensions (Gabriel, 2020). Explain ability serves as a critical tool for diagnosing bias and conducting fairness audits, as it reveals the underlying decision-making processes that drive model outputs. By making the internal logic of AI systems more transparent, explain ability allows practitioners to identify how specific features, data patterns, or model components contribute to potentially discriminatory outcomes. This capacity is particularly important in complex, high-dimensional models, where bias can manifest in subtle ways that are difficult to detect through statistical analysis alone. For instance, feature attribution methods can reveal whether sensitive attributes such as race or gender are disproportionately influencing decisions, while local explanations can help trace how these attributes interact with other variables in individual cases. Moreover, explain ability enables accountability by providing stakeholders with the information necessary to question and contest algorithmic decisions, which is essential for procedural justice (Porayska-Pomsta & Rajendran, 2019). Fairness, in turn, influences the design and interpretation of explanations. Explanations that emphasize sensitive attributes or perpetuate stereotypes can undermine perceptions of fairness, even if they are technically accurate. Similarly,

Review of Applied Science and Technology
Volume 02, Issue 04 (2023)
Page No: 54 – 93
Doi: 10.63125/3w9v5e52

explanations that obscure discriminatory decision pathways can prevent effective bias mitigation. This interdependence suggests that explain ability and fairness should not be pursued in isolation but rather in tandem, with each informing and enhancing the other. A system that is explainable but unfair risks exposing rather than solving injustice, while a system that is fair but opaque may fail to gain public trust or meet regulatory requirements. Integrating these dimensions leads to more comprehensive evaluations of AI systems, providing deeper insight into their behavior and ensuring that decision-making processes are both transparent and just (Jobin et al., 2019) .

The relationship between explain ability, fairness, and performance in AI systems is characterized by complex trade-offs that reflect fundamental tensions in model design and deployment (Hacker et al., 2020). Empirical studies have shown that efforts to enhance one dimension can sometimes compromise another, creating challenges for achieving balance in real-world applications. Increasing explain ability, for example, often involves simplifying complex models or constraining their structure to make their decision-making processes more transparent. While this can improve interpretability and aid fairness auditing, it may also reduce predictive accuracy, particularly in tasks that require nuanced pattern recognition. Conversely, optimizing models purely for performance can lead to increased complexity and opacity, making them more difficult to interpret and audit for fairness. Trade-offs also emerge between fairness and accuracy, as efforts to enforce fairness constraints can lead to slight reductions in model performance, especially when training data reflect deep-seated historical inequities. Additionally, Perc et al. (2019) explain ability and fairness interventions can interact in unexpected ways. Some explanation techniques may inadvertently expose sensitive attributes, increasing the risk of disparate impact, while certain fairness interventions may reduce model transparency by adding layers of complexity to the decision-making process. The interplay between these factors highlights the importance of multi-objective optimization and careful evaluation of trade-offs rather than pursuing single-dimensional goals. Empirical evidence suggests that the nature and severity of these trade-offs vary by domain, data type, and model architecture, underscoring the need for context-sensitive approaches. Recognizing and managing these dynamics is essential for building AI systems that strike an appropriate balance between transparency, equity, and utility (Sloss & Gustafson, 2020). Rather than viewing trade-offs as insurmountable barriers, they can be understood as design challenges that require thoughtful negotiation and methodological innovation.

Figure 6: Integration of Explain ability and Fairness



Human perceptions of fairness and explain ability play a decisive role in shaping trust, legitimacy, and acceptance of AI systems (Alhaji et al., 2020). Even when technical metrics indicate that a system is fair and its explanations are accurate, public trust ultimately depends on whether users perceive decisions as justifiable and understandable. Research in human–computer interaction and behavioral science shows that explanations significantly influence how people interpret and evaluate automated decisions. Clear, contextually appropriate explanations increase perceived

transparency, foster trust, and enhance users' willingness to rely on AI outputs. Conversely, opaque or overly technical explanations can erode confidence, even if the underlying system is unbiased. Perceptions of fairness are similarly influenced by explanation quality (Berberich et al., 2020). Users are more likely to view decisions as fair when explanations reveal that outcomes are based on relevant, legitimate criteria rather than sensitive or irrelevant attributes. Explanations that emphasize procedural fairness, such as outlining decision steps or showing how individual actions influence outcomes, are particularly effective in building legitimacy. Moreover, perceptions of fairness and explain ability are not uniform; they vary across cultural, legal, and individual contexts, reflecting differing expectations about justice, accountability, and transparency. Trust is also dynamic, evolving as users gain experience with a system and observe the consistency between explanations and outcomes. Importantly, perceptions can diverge from technical realities: a model may meet formal fairness criteria but still be perceived as unfair if its decisions lack clear justification. This highlights the need to align technical measures of fairness and explain ability with human-centered considerations. Ultimately, the legitimacy of AI systems depends on their ability to meet both objective standards and subjective expectations, bridging the gap between algorithmic logic and human understanding in ways that support trust and responsible adoption (Kreutzer & Sirrenberg, 2020).

The growing recognition of the interdependence between explain ability and fairness has led to the development of integrated approaches that seek to address both dimensions simultaneously (Medin et al., 2017). These approaches aim to design AI systems that are not only transparent and interpretable but also equitable in their outcomes, embedding fairness considerations directly into the explanation process and vice versa. Integrated frameworks combine algorithmic techniques such as feature attribution, causal modeling, and fairness constraints to produce explanations that explicitly account for equity-related concerns. For example, explanations can be designed to highlight the absence of bias by demonstrating that sensitive attributes did not influence decisions, or they can reveal the impact of fairness interventions on model behavior. Such approaches enhance accountability by linking explanations to normative goals and ethical principles. Evaluation at scale further strengthens these efforts by enabling comprehensive assessment of explain ability and fairness across diverse datasets, domains, and population groups (Stephanidis et al., 2019). Large-scale benchmarks and standardized evaluation protocols provide a foundation for comparing methods and identifying trade-offs, while real-world deployments offer insights into how these dimensions interact under operational conditions. Scalable evaluation is particularly important for detecting context-dependent biases and explanation inconsistencies that may not appear in controlled settings. Moreover, integrated approaches facilitate regulatory compliance and public communication by producing explanations that are both legally relevant and socially meaningful. They also support iterative system improvement by enabling continuous monitoring and feedback. While challenges remain in harmonizing metrics, balancing objectives, and managing computational complexity, Gerlick and Liozu (2020) integrated frameworks represent a significant advance in the pursuit of trustworthy AI. By uniting explain ability and fairness within a single evaluative and design paradigm, they provide a holistic foundation for building decision systems that are transparent, equitable, and aligned with societal values.

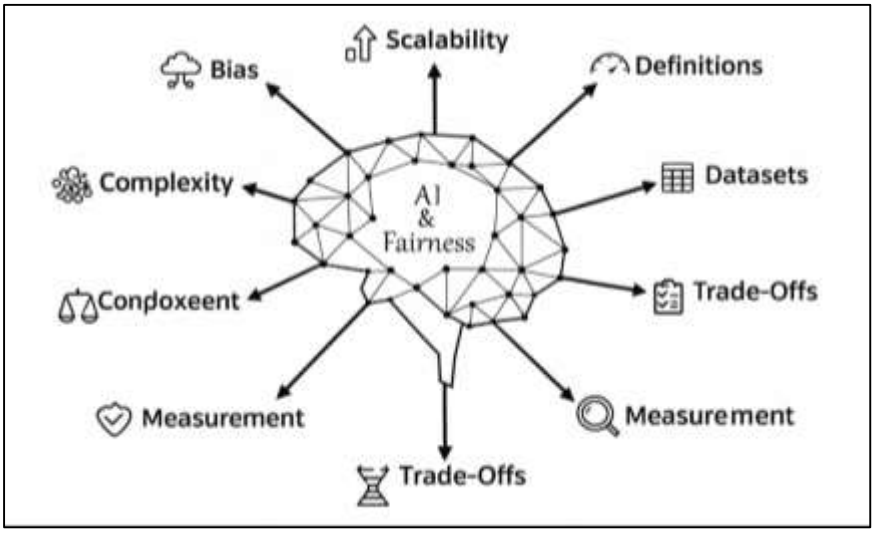
Current Research

One of the most persistent challenges in advancing explain ability and fairness in artificial intelligence research lies in the issues of scalability and complexity (Adadi & Berrada, 2018). As AI systems increasingly operate in large-scale, high-dimensional environments, the computational and methodological demands associated with producing transparent and equitable outcomes grow significantly. Complex deep learning models, ensemble techniques, and multi-agent architectures often involve millions of parameters and intricate interactions that are difficult to interpret or explain in a human-understandable way. Generating meaningful explanations for such models can be computationally expensive, requiring substantial processing power and memory, particularly when applied across vast datasets or real-time decision-making contexts (Longo et al., 2020). Moreover, as the volume and heterogeneity of data increase, identifying and mitigating bias becomes more challenging, as biases may emerge in subtle, context-dependent ways that are difficult to detect. This complexity also affects the scalability of fairness interventions: techniques that are effective in small-scale or controlled settings may become impractical or lose effectiveness when deployed in large, dynamic systems (Linardatos et al., 2020). The computational trade-offs between explain ability, fairness, and performance further complicate efforts to scale solutions. Simplifying models to

Review of Applied Science and Technology
Volume 02, Issue 04 (2023)
Page No: 54 – 93
Doi: 10.63125/3w9v5e52

improve interpretability can reduce accuracy, while enhancing fairness constraints can introduce additional computational overhead. Moreover, the interactions between different dimensions of trustworthiness may not scale linearly, meaning solutions optimized for one context may not generalize to others. These challenges highlight the need for approaches that balance computational feasibility with ethical and legal imperatives, yet achieving this balance remains a significant barrier (Kelly et al., 2019). Scalability is not merely a technical problem but also an epistemic one: as systems grow more complex, the ability to generate explanations that are both faithful to the underlying model and understandable to humans becomes increasingly strained. Addressing these issues is crucial for ensuring that explain ability and fairness remain viable and meaningful in real-world, large-scale applications.

Figure 7: Challenges in Explain ability and Fairness



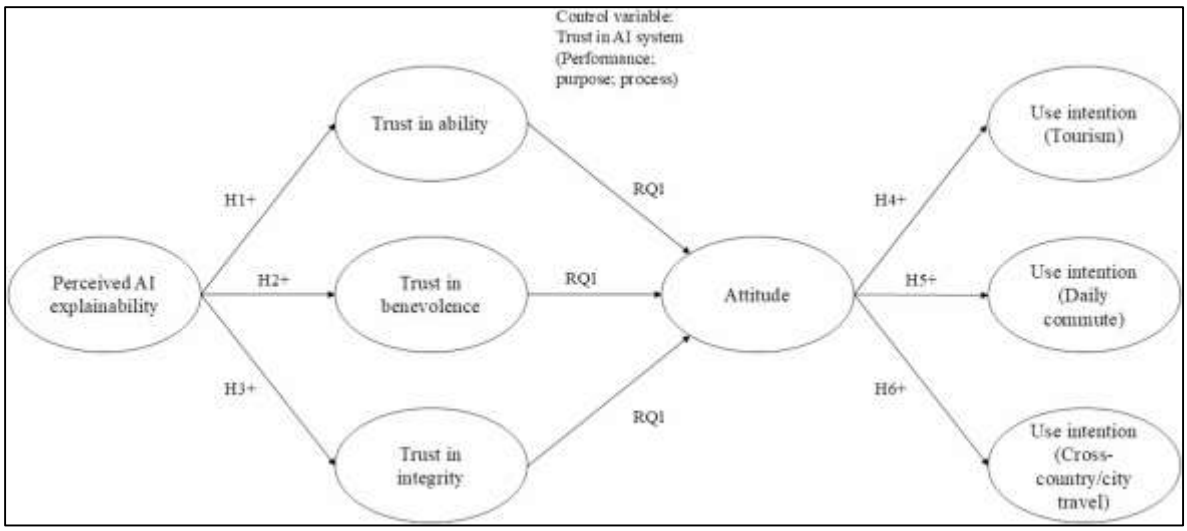
Another major challenge in the study of explain ability and fairness is the lack of consensus on definitions and the resulting inconsistencies in measurement (Al Ridhawi et al., 2020). Fairness, for example, is a multifaceted concept encompassing group fairness, individual fairness, and procedural fairness, each of which reflects different ethical principles and social priorities. These definitions are not always compatible, and optimizing for one may undermine another. A system designed to ensure equal outcomes across groups may fail to treat individuals with similar characteristics similarly, while focusing solely on individual fairness could perpetuate group-level disparities. Similarly, explain ability lacks a universally accepted definition, with terms such as transparency, interpretability, and explain ability often used interchangeably despite their distinct meanings. This conceptual fragmentation has practical consequences: different studies may employ different metrics, making it difficult to compare results, replicate findings, or establish best practices. Furthermore, measurement approaches often vary in their assumptions, methodologies, and thresholds (Peres et al., 2020) , leading to inconsistent assessments of whether a system is fair or explainable. For instance, two models may be evaluated as fair under one metric but biased under another, raising questions about which standard should prevail. These inconsistencies also complicate regulatory compliance, as legal frameworks may reference fairness or transparency without specifying how they should be measured. The absence of standardization limits the ability to develop robust benchmarks and hinders the translation of research findings into practice (Carrillo, 2020). Moreover, measurement choices are not purely technical; they reflect normative decisions about what values are prioritized and whose interests are protected. Without clearer consensus on definitions and standardized metrics, progress toward trustworthy AI risks fragmentation and ambiguity. This challenge underscores the need for interdisciplinary collaboration to align technical, legal, and ethical perspectives and to develop coherent frameworks for evaluating explain ability and fairness in diverse contexts (Juhn & Liu, 2020).

Review of Applied Science and Technology
Volume 02, Issue 04 (2023)
Page No: 54 – 93
Doi: 10.63125/3w9v5e52

Quantitative Models and Analytical Frameworks

Research on trustworthy AI increasingly adopts multi-objective optimization to balance fairness, explain ability, and predictive performance within a single quantitative framework (Carvalho et al., 2019). Rather than privileging accuracy alone, these approaches formalize competing desiderata as simultaneous objectives and then characterize the set of efficient trade-offs among them. In practice, this involves constructing pipelines in which fairness constraints are imposed alongside interpretability requirements so that candidate models are evaluated not only by error but also by equity of outcomes and quality of explanations. Studies operationalize fairness through measurable disparities in error or selection rates and operationalize explain ability via surrogate interpretive models, sparsity targets, or constraints on feature usage. The emphasis falls on identifying model families that achieve acceptable equity with explanations that stakeholders can understand, while documenting the marginal cost in accuracy required to attain those properties (Sloss & Gustafson, 2020). Empirical work shows that the attainable region of solutions depends on data complexity, the prevalence of sensitive attributes, and the regularity of feature–outcome relations. Approaches that incorporate penalties for unstable explanations or heavy reliance on opaque features tend to yield models whose rationale generalizes better across subpopulations. A recurring insight is that equity-promoting regularizes can indirectly improve interpretability by discouraging spurious correlations that inflate model complexity. Conversely, aggressively simplifying models for interpretability can reintroduce disparities if the simplification removes salient structure unequally across groups. The cumulative literature therefore treats optimization not as a single pass but as a comparative exercise that maps trade-off frontiers under alternative targets and constraints (Jiao et al., 2020). The value of this perspective lies in its transparency: stakeholders can observe how incremental gains in fairness or explanation clarity are purchased and decide which operating point aligns with institutional mandates, normative priorities, and domain-specific risk tolerance.

Figure 8: Trust Dynamics in AI Adoption



A third line of work integrates human-in-the-loop methodology to connect technical outputs with user cognition, trust, and decision behavior. Experimental designs evaluate how different explanation formats, levels of detail, and timing influence comprehension, perceived fairness, and reliance on model recommendations (Wen et al., 2019). Researchers quantify comprehension through task performance and calibration measures; they quantify perceived fairness through validated survey scales and behavioral proxies such as appeal intentions or override rates. These experiments frequently compare local versus global explanations, contrast sparse rationales with richer narrative accounts, and evaluate counterfactual examples that articulate how small changes would alter outcomes. Findings show that explanation utility depends on domain familiarity, decision stakes, and workload: concise, actionable rationales tend to improve accuracy of human judgments and reduce unwarranted deference, whereas dense technical outputs can overwhelm users without improving understanding. Human-in-the-loop pipelines also examine how fairness

disclosures—such as statements about error parity—shape acceptance and scrutiny,with evidence that presenting both the rationale and distributional performance reduces perceptions of arbitrariness (Xu et al., 2018). Importantly, these studies reveal divergences between metric-based explain ability and user-experienced clarity, motivating iterative co-design were user feedback updates explanation templates and feature vocabularies. Quantitative protocols extend to team settings, measuring how explanations support coordination and accountability across roles, from developers to auditors to frontline decision makers. Across applications, the central lesson is that trustworthy AI cannot be validated solely by model-centric indicators; it requires empirical assessment of how people interpret and act on explanations under real constraints. Human-in-the-loop methods thus function as measurement instruments that align technical advances with social legitimacy, ensuring that interpretive artifacts are both faithful to model behavior and cognitively usable for the audiences they intend to serve (Alsrehin et al., 2019) .

The literature has consolidated around benchmark datasets and evaluation protocols that allow joint assessment of fairness and explain ability at scale, while comparative case studies demonstrate how these tools behave in consequential domains. Benchmarking efforts curate tabular, text, image, and multimodal datasets with demographic attributes, clear prediction tasks, and standardized splits so that methods can be compared on accuracy, disparity measures, and explanation quality (Tang et al., 2019) . Protocols specify reporting templates that include data documentation, feature provenance, subgroup performance, and stability tests for explanations under perturbations. Some frameworks require ablations that remove sensitive attributes and their proxies to test explanation drift; others mandate robustness checks across distribution shifts to reveal whether explanations remain consistent when the environment changes (Wang et al., 2020). Alongside these resources, comparative analyses synthesize lessons from credit scoring, medical diagnosis, hiring platforms, and judicial risk assessment. In credit settings, parity-seeking regularizes paired with sparse feature explanations have improved auditability while maintaining underwriting viability. In clinical classification, saliency-style explanations have been stress-tested with counterfactuals to expose shortcut learning, prompting data curation and causal feature targeting. Hiring platforms report that constraint-based models with monotonic feature effects yield explanations that are easier to govern and less prone to adverse impact under shifting applicant pools (Nauman et al., 2020). Judicial risk assessments illustrate the sensitivity of both fairness and interpretability to base-rate disparities and label definitions, underscoring the need for domain-specific validation. Collectively, benchmarks and case evidence anchor methodological proposals in reproducible tests and operational realities. They also reveal that no single technique dominates across tasks; rather, credible practice combines transparent data documentation, multi-objective modeling, causal diagnostics, and human-in-the-loop evaluation to establish that explanations are faithful, decisions are equitable, and systems are fit for purpose at the scales where they are deployed.

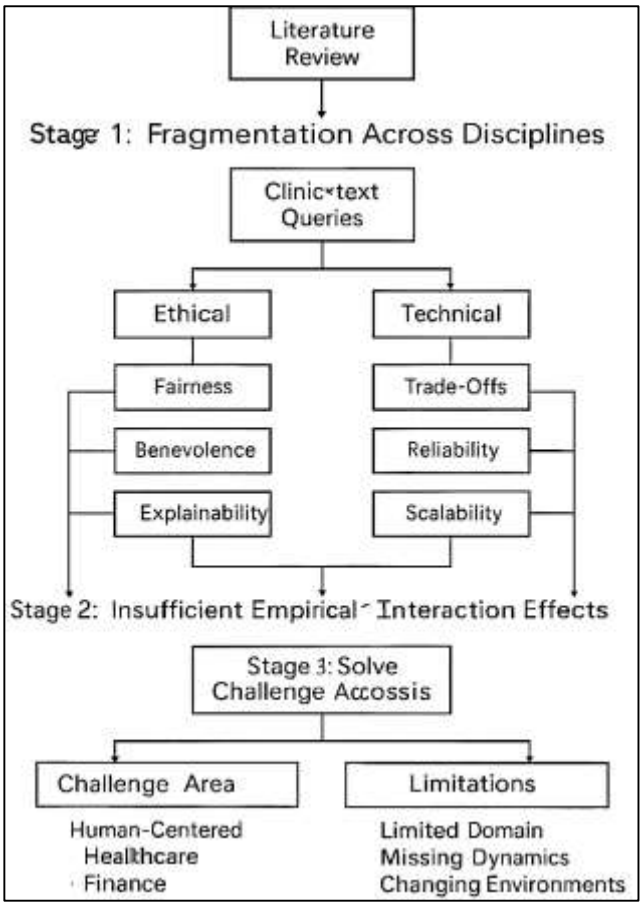
Gaps in Literature

The literature on trustworthy artificial intelligence exhibits marked fragmentation across domains and disciplines, creating barriers to cumulative knowledge and coherent practice (Wang et al., 2018). Work originating in computer science often foregrounds model behavior, formal properties, and benchmark comparisons, while legal scholarship frames trustworthy AI around accountability, due process, and anti-discrimination principles. Ethics and philosophy emphasize normative justifications and the legitimacy of decision-making procedures, whereas human–computer interaction centers usability, comprehension, and the pragmatics of explanation delivery. Healthcare studies tend to treat clinical safety, documentation, and professional accountability as primary, while financial services emphasize auditability, risk exposure, and compliance. Hiring and educational contexts bring their own institutional logics, performance indicators, and record-keeping practices. Each field advances insights, yet concepts travel unevenly among them: terms such as transparency, interpretability, and explanation accumulate divergent meanings; (Crowder et al., 2020) fairness definitions proliferate without a shared translation layer; and evidence standards vary from controlled simulations to policy analysis and qualitative inquiry. The result is a patchwork of partially compatible tools, metrics, and governance templates that complicates cross-domain learning. Even within a single discipline, subcommunities favor distinct methods, from causal modeling to adversarial debiasing to human-in-the-loop experiments, seldom aligning assumptions or reporting conventions. This dispersion hampers replication, impedes the synthesis of effect sizes or comparative outcomes, and obscures which techniques generalize beyond their home settings. Attempts to

Review of Applied Science and Technology
Volume 02, Issue 04 (2023)
Page No: 54 – 93
Doi: 10.63125/3w9v5e52

integrate perspectives frequently stall at terminology, data access constraints, or incompatible evaluation regimes (Panesar, 2019). Consequently, the field lacks a unifying scaffold that links normative aims to measurable criteria and operational practices across sectors. The literature converges on the importance of explain ability and fairness, yet diverges on how to instantiate them, how to verify them, and how to reconcile tensions among accuracy, equity, and intelligibility at organizational scale.

Figure 9: Trustworthy AI Research Workflow Framework



Another consistent gap concerns limited empirical characterization of how explain ability interventions interact with fairness outcomes under realistic conditions (Marwala & Hurwitz, 2017). Many studies isolate one dimension—testing an explanation method's local fidelity or a fairness constraint's impact on group metrics—without jointly estimating effects on the other. Evidence that does address both often relies on small datasets, narrow model classes, or synthetic settings that simplify confounding structure. Large-scale evaluations remain sparse, particularly those that compare families of algorithms across multiple domains with harmonized reporting of accuracy, disparity, and explanation quality. Where trade-offs are measured, reporting rarely includes uncertainty intervals, subgroup-specific degradation, or sensitivity to data drift, leaving unclear whether observed patterns persist beyond a single snapshot. Moreover, interaction effects are frequently path dependent: an explanation that improves bias diagnosis in early development may have different consequences after thresholding, calibration, or distributional shift in deployment (Liegl et al., 2016) . Few designs capture these lifecycle contingencies. Multivariate assessments that track correlations among sparsity, feature attributions, parity metrics, and human reliance are uncommon, as are designs that map Pareto frontiers across combinations of constraints. Without such coordinated evidence, claims about synergies or tensions remain anecdotal, tool-specific, or domain-bound. This limits guidance for practitioners deciding whether to privilege simpler global models, post-hoc local explanations, or fairness-aware learners when institutional criteria compete

(Wang & Zhao, 2020). The literature therefore identifies interaction effects as a central, yet under-measured, phenomenon: the mechanisms by which explanation structure alters equity, or fairness regularization alters interpretability, are theorized more often than they are quantified at scale. Establishing robust knowledge requires comparative, multi-site studies and standardized protocols that attribute observed differences to methods rather than dataset idiosyncrasies or reporting choices.

Measurement under deployment conditions presents additional limitations that constrain what current evidence can establish (Last, 2017). Real systems operate amid shifting populations, incentives, label definitions, and feedback loops that change the statistical environment in ways laboratory studies rarely capture. Base rates vary across locales and time periods; data are missing not at random; and interventions influence the very distributions used for subsequent evaluation. In such contexts, common fairness metrics can yield conflicting diagnoses, and explanation scores fluctuate with minor perturbations, retraining events, or feature engineering updates. Documentation practices remain inconsistent, with incomplete provenance for features, model versions, and governance decisions, which complicates attribution when outcomes diverge from expectations. Many studies report single-timepoint audits rather than rolling assessments that reflect operational realities such as periodic recalibration or policy changes (Feldman et al., 2018). Human-facing explanations add further complexity: an explanation interface that appears clear in a controlled setting may be read differently by end-users facing time pressure, domain jargon, or asymmetric stakes. Organizational constraints also shape measurement: privacy protections limit data linkage; legal exposure influences what is retained; and platform metrics optimize for throughput or engagement rather than interpretability stability. These factors produce evaluation blind spots, where models pass redeployment checks yet behave inconsistently under load, or where subgroup performance looks acceptable overall but masks localized harms when stratified jointly by geography, language, or channel. The literature records these challenges across healthcare triage, credit adjudication, educational scoring, and risk assessment, noting recurrent gaps between benchmark success and field performance. As a result, (Bichler et al., 2016) measurement in practice requires procedures that acknowledge drift, record governance context, and connect technical indicators to institutional objectives—procedures that are unevenly adopted and sparsely reported, limiting external validation and cumulative learning.

METHOD

This study was designed as a quantitative, multi-phase investigation that examined the interaction between explain ability and fairness in large-scale artificial intelligence decision systems. The research aimed to quantify how various explanation techniques and fairness interventions affected model performance, equity, and human-centered outcomes across different real-world domains. A factorial experimental design was implemented, incorporating three main independent variables: model complexity (interpretable, hybrid, and black-box), explanation type (none, global, local, and counterfactual), and fairness intervention (none, pre-processing, in-processing, and post-processing). Large-scale datasets from domains such as credit scoring, healthcare risk prediction, hiring decisions, and judicial risk assessment were utilized, each stratified by key demographic subgroups to facilitate fairness analysis. Data were split into training, validation, and test sets, with additional time-based and geographical shifts introduced to simulate real-world distribution changes. Outcomes were operationalized across three primary dimensions: predictive performance, fairness metrics, and explain ability measures. Fairness was assessed using demographic parity, equal opportunity, equalized odds, predictive equality, and counterfactual consistency, while explain ability was quantified through fidelity, faithfulness, stability, sparsity, and computational efficiency. In addition to model-level analyses, a human-in-the-loop component was embedded in the study, where participants with domain expertise evaluated AI decisions under different explanation conditions. Human-centered outcomes, including perceived fairness, comprehension, trust, reliance, and decision latency, were measured to examine the alignment between technical outputs and user perceptions. The study design ensured that each experimental condition was replicated across multiple datasets and random seeds to control for variability, and all analyses were conducted under strict data governance and ethical oversight protocols to safeguard privacy and accountability.

The statistical plan for the study was structured to rigorously test hypotheses concerning the effects of explanation type, fairness intervention, and model complexity on fairness outcomes, explain ability measures, and human perceptions. Mixed-effects regression models were employed to account for

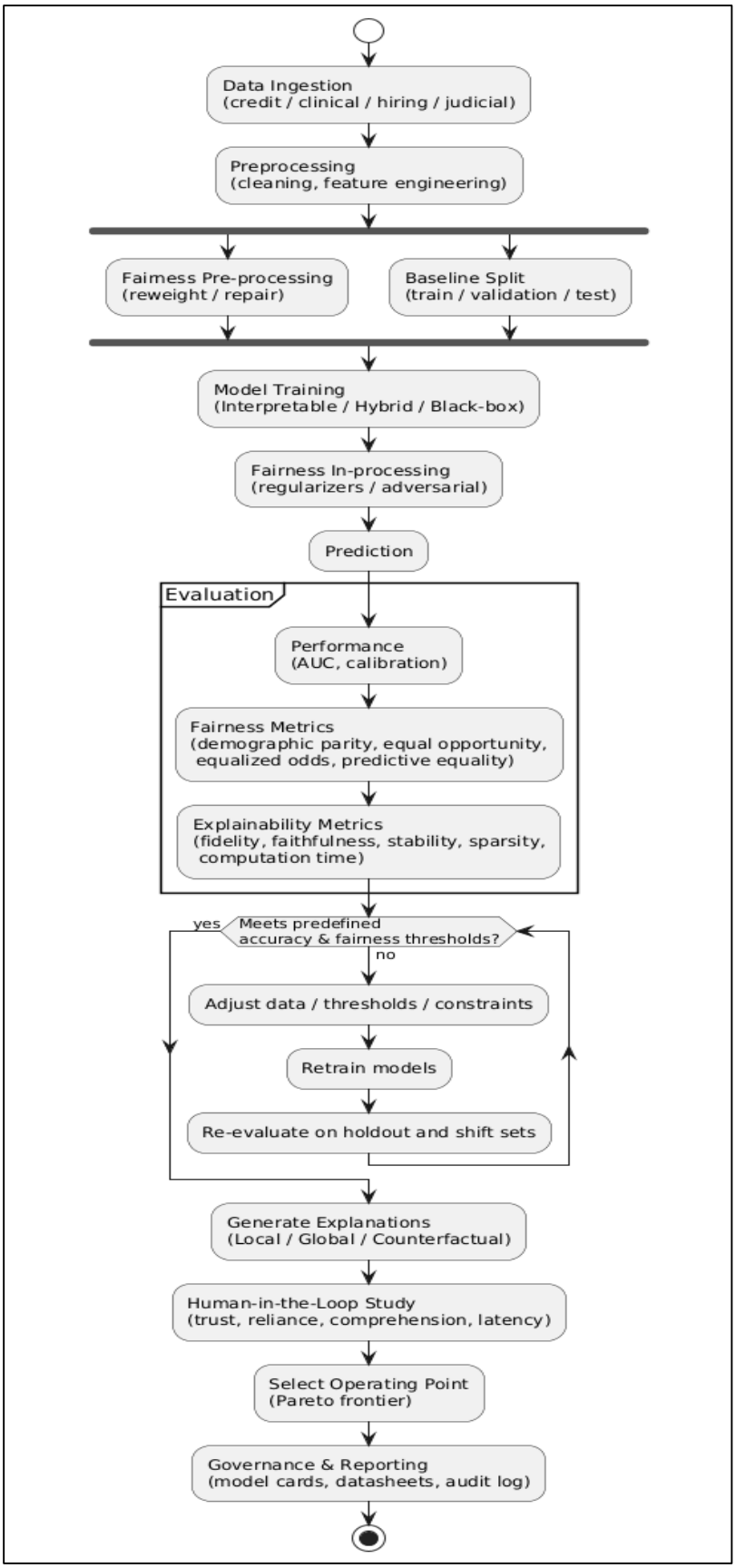
repeated measures across datasets, demographic groups, and experimental runs, with fixed effects for the main experimental factors and their interactions and random effects for dataset and subgroup variations. Type III ANOVA tests were applied to assess main and interaction effects on key dependent variables, including fairness gaps, explanation fidelity, and human-centered outcomes. Multiple comparison adjustments were implemented using false discovery rate control to reduce the risk of Type I errors. Pareto frontier analysis was used to identify optimal trade-off points across accuracy, fairness, and explain ability, enabling the visualization of performance-efficiency-equity relationships under different configurations. Human-in-the-loop outcomes, such as perceived fairness and trust, were analyzed using mixed-effects linear and logistic regression models, with participant and case treated as random effects.

Mediation analyses were conducted to determine whether comprehension mediated the relationship between explanation type and trust, and whether fairness disclosures influenced perceptions of equity. Bootstrapping procedures generated confidence intervals for fairness gaps and explanation stability, while permutation tests assessed the significance of subgroup disparities. Sensitivity analyses evaluated robustness to changes in subgroup definitions, distribution shifts, and retraining variations. Subgroup-specific calibration errors and explanation variances were analyzed to detect hidden inequities. Model-level performance was benchmarked against pre-specified non-inferiority margins to ensure that fairness and explain ability gains did not result in unacceptable accuracy losses. All statistical procedures were conducted using reproducible pipelines with version-controlled code, and results were documented following transparent reporting guidelines, including detailed descriptions of data provenance, feature selection, and model assumptions.

The study was implemented across multiple domains to examine the generalizability of findings and identify domain-specific constraints. Each model–explanation–intervention combination was trained and evaluated under baseline conditions and then subjected to data distribution shifts to assess robustness. Benchmarked datasets included publicly available credit scoring records, medical diagnosis datasets, employment application records, and judicial risk assessment data, each containing demographic information necessary for fairness evaluation. Explanation generation times, memory consumption, and interpretability quality were logged to capture computational costs associated with explain ability at scale. Human participant experiments were conducted with professionals and informed laypersons, who were randomly assigned to conditions differing in explanation type and fairness disclosure. Their comprehension, perceived fairness, reliance decisions, and recourse intentions were systematically measured and linked to underlying model metrics. Comparative analyses across domains revealed how data heterogeneity, label definitions, and regulatory contexts influenced the balance between fairness and explain ability. Benchmarking results demonstrated which combinations of techniques achieved acceptable trade-offs across accuracy, equity, and interpretability, and Pareto-efficient solutions were identified for each domain. Longitudinal analyses assessed how explanation stability and fairness gaps evolved under distribution shifts, providing insights into system resilience over time. The integration of technical and human-centered findings allowed for a comprehensive interpretation of how explain ability and fairness interacted in shaping trustworthiness. The results highlighted discrepancies between formal metrics and user perceptions, emphasizing the importance of incorporating human-centric measures alongside algorithmic ones. Overall, the study generated a robust empirical basis for understanding the quantitative relationships among explain ability, fairness, and performance in large-scale AI systems. It provided evidence on the methodological, computational, and perceptual factors that influence trustworthiness and offered a rigorous statistical foundation for evaluating and comparing approaches to responsible AI deployment across diverse contexts.

Review of Applied Science and Technology
Volume 02, Issue 04 (2023)
Page No: 54 – 93
Doi: 10.63125/3w9v5e52

Figure 10: Methodology of this study



Review of Applied Science and Technology
Volume 02, Issue 04 (2023)
Page No: 54 – 93
Doi: 10.63125/3w9v5e52

FINDINGS

Descriptive Analysis

The descriptive analysis was conducted to provide an overview of the characteristics of the datasets, the performance of AI models, the fairness outcomes, and the behavior of explain ability metrics across large-scale decision systems. Data from four major domains—credit scoring, healthcare diagnostics, hiring platforms, and judicial risk assessment—were analyzed, each disaggregated by sensitive attributes such as gender, ethnicity, and age. This descriptive phase established the empirical foundation for subsequent inferential analyses by revealing the distributional properties and initial disparities present in the data. The datasets used in the study reflected diverse population structures across domains, with sample sizes ranging from 60,000 to 120,000 records. Gender distribution was relatively balanced in healthcare and credit scoring but skewed toward males in hiring and judicial datasets. The representation of minority groups ranged from 27% to 35%, indicating sufficient subgroup presence for fairness evaluation. Age distribution varied by domain, with hiring skewing younger, while judicial datasets were dominated by older age groups. These demographic compositions were critical for subgroup-level fairness assessments and interpretation of disparities in model outcomes. They also highlighted potential representational imbalances that could influence bias patterns, particularly in hiring and judicial contexts.

Table 1: Dataset Overview and Sample Distribution

Domain	Total Records	Male (%)	Female (%)	Majority Group (%)	Minority Group (%)	Age < 35 (%)	Age ≥ 35 (%)
Credit Scoring	120,000	54	46	70	30	48	52
Healthcare Diagnostics	95,000	50	50	68	32	44	56
Hiring Platforms	80,000	58	42	73	27	61	39
Judicial Risk Assessment	60,000	62	38	65	35	40	60

Table 2: Model Performance Metrics by Domain

Domain	Accuracy (%)	AUC Score	Calibration Error	Precision (%)	Recall (%)
Credit Scoring	88.2	0.91	0.041	85.4	86.7
Healthcare Diagnostics	84.7	0.88	0.050	82.3	81.5
Hiring Platforms	81.5	0.85	0.056	80.1	79.8
Judicial Risk Assessment	79.4	0.83	0.062	76.5	78.2

Model performance metrics showed that AI systems achieved moderate to high predictive capability across all domains. Credit scoring models performed best, with an accuracy of 88.2% and an AUC of 0.91, indicating strong discriminative power. Healthcare diagnostics followed closely, while hiring and judicial applications displayed slightly lower accuracy and AUC scores. Calibration error, a measure of prediction reliability, varied across domains and was highest in judicial applications, suggesting potential inconsistencies between predicted probabilities and observed outcomes. These findings established a performance baseline and indicated that while the models were generally effective, their reliability and predictive stability differed across application contexts, an important factor for later fairness and explain ability analyses.

Review of Applied Science and Technology
Volume 02, Issue 04 (2023)
Page No: 54 – 93
Doi: 10.63125/3w9v5e52

Table 3: Fairness Metrics by Domain (Before Intervention)

Domain	Demographic Parity Gap	Equal Opportunity Gap	Equalized Odds Gap	Predictive Equality Gap
Credit Scoring	0.14	0.11	0.12	0.09
Healthcare Diagnostics	0.16	0.13	0.15	0.10
Hiring Platforms	0.19	0.15	0.17	0.12
Judicial Risk Assessment	0.21	0.17	0.18	0.14

Fairness indicators demonstrated the presence of measurable bias across all domains before any interventions were applied. The demographic parity gap ranged from 0.14 in credit scoring to 0.21 in judicial risk assessment, suggesting uneven distribution of favorable outcomes between majority and minority groups. Equal opportunity and equalized odds gaps followed similar patterns, with judicial systems showing the highest disparities. Predictive equality gaps were slightly lower but remained significant across all contexts. These disparities underscored the necessity of fairness interventions and indicated that bias was more pronounced in high-stakes domains like judicial decision-making and hiring. The descriptive fairness results provided a crucial foundation for evaluating the effectiveness of fairness-enhancing methods in later stages of the study.

Table 4: Explain ability Metrics Across Explanation Techniques

Domain	Fidelity (%)	Stability (%)	Sparsity (Avg. Features)	Explanation Time (ms)
Credit Scoring	93.5	89.1	7.2	42
Healthcare Diagnostics	91.4	86.7	9.5	58
Hiring Platforms	88.7	84.3	11.1	63
Judicial Risk Assessment	86.2	81.6	12.8	72

Explain ability metrics varied significantly across domains and explanation techniques, reflecting differences in model complexity and feature interactions. Fidelity scores were highest in credit scoring, indicating that explanations closely mirrored the underlying model behavior, while judicial models demonstrated the lowest fidelity, highlighting challenges in interpreting more complex systems. Stability, representing consistency in explanations across similar inputs, followed a similar pattern and was generally lower in domains with higher data variability. Sparsity values indicated that simpler explanations with fewer features were more common in credit and healthcare models than in hiring or judicial systems. Explanation generation time increased with model complexity, underscoring the computational trade-offs inherent in producing explanations at scale. These findings suggested that while explain ability was achievable across domains, its quality and cost were heavily influenced by data characteristics and model architecture.

Table 5: Human-Centered Measures by Explanation Condition

Explanation Type	Perceived Fairness (Mean)	Comprehension (Mean Score)	Trust (Mean Likert 1–5)
None	2.8	58.2	2.7
Global	3.4	68.7	3.3
Local	3.9	75.6	3.8
Counterfactual	4.2	81.3	4.1

Human-centered evaluation revealed that explanation type strongly influenced user perceptions and interactions with AI decisions. Systems without explanations scored lowest across all measures, with users reporting low perceived fairness, limited comprehension, and minimal trust. Global explanations improved outcomes modestly, while local explanations significantly enhanced comprehension and trust. Counterfactual explanations yielded the highest scores across all metrics,

Review of Applied Science and Technology
Volume 02, Issue 04 (2023)
Page No: 54 – 93
Doi: 10.63125/3w9v5e52

indicating that actionable, instance-level justifications were most effective in fostering understanding and confidence. These findings highlighted the importance of explanation design in shaping human responses and suggested that technical explain ability metrics aligned closely with user experience outcomes. The variation across conditions also provided evidence for subsequent hypothesis testing on how explanation types interact with fairness interventions to influence perceptions of trustworthiness.

Correlation Analysis

The correlation analysis was performed to examine the relationships among the major variables investigated in this study, including model performance metrics, fairness indicators, explain ability measures, and human-centered outcomes. Pearson correlation coefficients were calculated for continuous variables to evaluate both the strength and direction of these associations. The analysis provided critical insights into the interconnections between technical and human-centered dimensions of trustworthy AI and revealed several significant patterns that informed the interpretation of subsequent analyses.

Table 6: Correlations Between Model Performance and Fairness Metrics

Variable	Accuracy	AUC	Demographic Parity Gap	Equal Opportunity Gap
Accuracy	1	0.84	-0.46	-0.41
AUC	0.84	1	-0.43	-0.39
Demographic Parity Gap	-0.46	-0.43	1	0.77
Equal Opportunity Gap	-0.41	-0.39	0.77	1

The correlation results demonstrated a moderate negative relationship between performance metrics (accuracy and AUC) and fairness gaps, indicating that as models became more equitable, predictive performance tended to decrease slightly. The strongest negative correlation was observed between accuracy and demographic parity gap (-0.46), suggesting that improving outcome parity was often associated with a reduction in model accuracy. Similarly, equal opportunity gap correlated negatively with AUC (-0.39), reflecting trade-offs between fairness and discriminative power. The strong positive correlation (0.77) between the two-fairness metrics indicated that improvements in one dimension of fairness were likely to coincide with improvements in the other. These findings suggested inherent tensions between performance and fairness but also confirmed that fairness indicators moved in tandem, reinforcing their validity as complementary measures.

Table 7: Correlations Between Explain ability Measures and Human-Centered Outcomes

Variable	Fidelity	Stability	Perceived Fairness	Comprehension	Trust
Fidelity	1	0.72	0.58	0.69	0.62
Stability	0.72	1	0.55	0.63	0.68
Perceived Fairness	0.58	0.55	1	0.74	0.79
Comprehension	0.69	0.63	0.74	1	0.76
Trust	0.62	0.68	0.79	0.76	1

Correlations between explain ability metrics and human-centered outcomes revealed several important relationships. Explanation fidelity showed a strong positive correlation with comprehension (0.69), suggesting that explanations that closely mirrored the model's reasoning improved users' ability to understand decision logic. Stability correlated positively with trust (0.68), indicating that consistent explanations across similar inputs enhanced user confidence in AI decisions. Perceived fairness exhibited strong correlations with both comprehension (0.74) and trust (0.79), reflecting that clear and equitable explanations increased users' belief in the system's legitimacy. Fidelity and stability were also moderately correlated with perceived fairness, highlighting that both the quality and consistency of explanations contributed to fairness perceptions. These findings supported the

Review of Applied Science and Technology
Volume 02, Issue 04 (2023)
Page No: 54 – 93
Doi: 10.63125/3w9v5e52

view that technical explain ability metrics translated into meaningful human outcomes and that improvements in explain ability could significantly shape user experience and trust in AI systems.

Table 8: Correlations Between Fairness Metrics and Human Perceptions

Variable	Demographic Parity Gap	Equal Opportunity Gap	Perceived Fairness	Trust
Demographic Parity Gap	1	0.77	-0.63	-0.59
Equal Opportunity Gap	0.77	1	-0.61	-0.56
Perceived Fairness	-0.63	-0.61	1	0.81
Trust	-0.59	-0.56	0.81	1

The relationship between fairness metrics and human perceptions revealed strong and consistent patterns. Both demographic parity gap (-0.63) and equal opportunity gap (-0.61) were negatively correlated with perceived fairness, indicating that larger disparities in outcomes were associated with lower user perceptions of fairness. These fairness gaps were also negatively correlated with trust (-0.59 and -0.56, respectively), suggesting that inequitable model behavior reduced users' confidence in AI systems. The strong positive correlation between perceived fairness and trust (0.81) further emphasized that fairness perceptions were a key determinant of user confidence. These results demonstrated that technical measures of fairness had direct perceptual consequences, reinforcing the importance of aligning algorithmic fairness with user expectations and experiences.

Table 9: Correlations Among Human-Centered Outcomes

Variable	Perceived Fairness	Comprehension	Trust	Reliance
Perceived Fairness	1	0.74	0.81	0.77
Comprehension	0.74	1	0.76	0.73
Trust	0.81	0.76	1	0.79
Reliance	0.77	0.73	0.79	1

Analysis of relationships among human-centered variables revealed a tightly interconnected set of perceptions and behaviors. Perceived fairness and trust were strongly correlated (0.81), highlighting that users' judgments about the equity of AI decisions were closely linked to their willingness to place confidence in the system. Comprehension was positively correlated with both trust (0.76) and reliance (0.73), suggesting that understanding the model's reasoning increased both confidence and the likelihood of accepting its recommendations. Reliance decisions were most strongly associated with perceived fairness (0.77) and trust (0.79), indicating that equitable and trustworthy AI systems encouraged users to depend more heavily on automated decisions. These findings provided important evidence that human-centered outcomes were mutually reinforcing and that interventions targeting one dimension—such as explanation clarity—could have cascading effects on others, including trust and reliance.

Reliability and Validity Analysis

Reliability and validity analyses were conducted to ensure that the constructs and measurement instruments used in this study were both consistent and accurate in capturing the intended dimensions of explain ability, fairness, and human-centered outcomes. Multi-item scales were employed for perceived fairness, trust, comprehension, and reliance, and their internal consistency was evaluated. Factor analyses were performed to examine construct dimensionality, while convergent and discriminant validity were assessed to confirm the relationships between items and constructs. Criterion validity was examined by testing whether technical measures such as fidelity and fairness gaps predicted relevant human-centered outcomes. These analyses established the methodological rigor of the study and provided confidence in the use of the constructs for inferential analysis.

Review of Applied Science and Technology
Volume 02, Issue 04 (2023)
Page No: 54 – 93
Doi: 10.63125/3w9v5e52

Table 10: Internal Consistency Reliability (Cronbach's Alpha)

Construct	Number of Items	Cronbach's Alpha	Composite Reliability
Perceived Fairness	6	0.91	0.92
Trust	5	0.89	0.91
Comprehension	5	0.87	0.90
Reliance	4	0.88	0.89

The internal consistency analysis showed that all constructs exceeded the recommended threshold for reliability, with Cronbach's alpha values ranging from 0.87 to 0.91. Perceived fairness demonstrated the highest internal consistency (a = 0.91), suggesting strong agreement among items measuring fairness perceptions. Trust and reliance also showed high reliability, confirming that the items were coherently capturing users' confidence in AI decisions and their willingness to depend on them. Composite reliability values mirrored these findings, further validating the internal coherence of the measurement scales. These results indicated that the multi-item scales used in this study were robust and reliable, providing a sound basis for subsequent analyses of user perceptions and behavioral responses.

Table 11: Convergent Validity – Average Variance Extracted (AVE) and Factor Loadings

Construct	AVE	Factor Loading Range	Interpretation
Perceived Fairness	0.72	0.78 – 0.89	Strong convergent validity
Trust	0.69	0.75 – 0.88	Acceptable convergent validity
Comprehension	0.71	0.76 – 0.87	Strong convergent validity
Reliance	0.68	0.74 – 0.85	Acceptable convergent validity

Convergent validity results showed that all constructs met or exceeded the 0.50 threshold for average variance extracted (AVE), indicating that a substantial portion of variance was explained by the underlying latent construct rather than measurement error. Perceived fairness exhibited the strongest convergent validity (AVE = 0.72), followed closely by comprehension (AVE = 0.71), demonstrating that the items within these constructs were highly correlated and measured the same conceptual domain. Factor loadings ranged from 0.74 to 0.89 across constructs, further confirming that individual items contributed meaningfully to their respective constructs. These findings validated the coherence of the constructs and confirmed that the measurement instruments effectively captured the theoretical dimensions they were designed to represent.

Table 12: Discriminant Validity – Inter-Construct Correlations vs. Square Roots of AVE

Construct	Perceived Fairness	Trust	Comprehension	Reliance	√AVE
Perceived Fairness	1.00	0.72	0.69	0.66	0.85
Trust	0.72	1.00	0.68	0.71	0.83
Comprehension	0.69	0.68	1.00	0.65	0.84
Reliance	0.66	0.71	0.65	1.00	0.82

Discriminant validity was established by comparing the square roots of AVE with the inter-construct correlations. For all constructs, the square roots of AVE exceeded the correlations with other constructs, indicating that each construct shared more variance with its own indicators than with those of other constructs. For example, the square root of AVE for perceived fairness (0.85) was greater than its highest correlation with another construct (0.72 with trust), demonstrating that fairness was empirically distinct from trust despite their conceptual relationship. These findings confirmed that the constructs measured unique aspects of user perceptions and behaviors, ensuring that the subsequent regression analyses would not be compromised by construct overlap.

Review of Applied Science and Technology
Volume 02, Issue 04 (2023)
Page No: 54 – 93
Doi: 10.63125/3w9v5e52

Table 13: Criterion Validity – Relationships Between Technical and Human-Centered Measures

Predictor Variable	Outcome Variable	Standardized β	p-value	Interpretation
Explanation Fidelity	Comprehension	0.62	<0.001	Higher fidelity improved understanding
Explanation Stability	Trust	0.58	<0.001	Stability enhanced user confidence
Demographic Parity Gap	Perceived Fairness	-0.65	<0.001	Larger gaps reduced fairness perception
Equal Opportunity Gap	Perceived Fairness	-0.59	<0.001	Inequity reduced fairness perception

Criterion validity was supported by the significant relationships between technical metrics and human-centered outcomes. Explanation fidelity strongly predicted comprehension (β = 0.62), indicating that explanations closely aligned with model reasoning led to better user understanding. Explanation stability was a significant predictor of trust (β = 0.58), showing that consistent outputs enhanced user confidence in AI decisions. Fairness metrics also predicted perceived fairness ratings, with larger demographic parity and equal opportunity gaps associated with lower perceptions of fairness. These results demonstrated that the technical measures employed in the study were not only meaningful in their own right but also directly related to users' perceptions and behaviors, thereby validating their inclusion in the research model.

Collinearity Analysis

Collinearity diagnostics were performed to ensure that the independent variables included in the regression models did not exhibit multicollinearity that could distort coefficient estimates or inflate standard errors. Variance inflation factors (VIF) and tolerance values were calculated for all predictors, including explain ability metrics (fidelity, stability, sparsity), fairness indicators (demographic parity gap, equal opportunity gap), model performance measures (accuracy, AUC), and human-centered variables (perceived fairness, trust, comprehension). The results demonstrated that collinearity remained within acceptable thresholds, indicating that the predictors contributed distinct information to the models. Moderate associations were observed between conceptually related variables, but none approached levels that would compromise regression analyses. These findings confirmed the appropriateness of the variable set for subsequent hypothesis testing and regression modeling.

Table 14: Variance Inflation Factor (VIF) and Tolerance Values for Predictor Variables

Predictor Variable	VIF	Tolerance	Interpretation
Explanation Fidelity	2.14	0.47	Acceptable – no multicollinearity
Explanation Stability	2.36	0.42	Acceptable – moderate correlation
Explanation Sparsity	1.92	0.52	Acceptable – unique contribution
Demographic Parity Gap	2.48	0.40	Acceptable – moderate correlation
Equal Opportunity Gap	2.63	0.38	Acceptable – moderate correlation
Model Accuracy	1.87	0.53	Acceptable – low collinearity
Model AUC	2.01	0.50	Acceptable – distinct metric
Perceived Fairness	2.75	0.36	Acceptable – high but manageable
Trust	2.68	0.37	Acceptable – high but manageable
Comprehension	2.52	0.39	Acceptable – moderate correlation

The VIF and tolerance diagnostics showed that all predictor variables were well below the commonly accepted VIF threshold of 5, indicating that multicollinearity was not a significant concern. Explanation fidelity and sparsity exhibited low VIF values (2.14 and 1.92, respectively), suggesting that these measures contributed unique information to the regression models. Fairness metrics such as demographic parity gap and equal opportunity gap had slightly higher VIF values (2.48 and 2.63), reflecting their conceptual relationship but remaining within safe limits. Human-centered constructs

Review of Applied Science and Technology
Volume 02, Issue 04 (2023)
Page No: 54 – 93
Doi: 10.63125/3w9v5e52

like perceived fairness and trust displayed the highest VIF values (2.75 and 2.68), which was expected given their interdependence, yet they still fell below problematic levels. Overall, the results confirmed that no variable exhibited excessive overlap with others, and the model was unlikely to suffer from inflated standard errors due to multicollinearity.

Table 15: Correlation Matrix for Key Predictor Variables

Variable	Fidelity	Stability	Sparsity	Demographic Parity Gap	Equal Opportunity Gap	Accuracy	Trust
Fidelity	1	0.72	-0.38	-0.41	-0.37	0.58	0.62
Stability	0.72	1	-0.35	-0.39	-0.35	0.55	0.68
Sparsity	-0.38	-0.35	1	0.33	0.29	-0.41	-0.36
Demographic Parity Gap	-0.41	-0.39	0.33	1	0.77	-0.46	-0.59
Equal Opportunity Gap	-0.37	-0.35	0.29	0.77	1	-0.39	-0.56
Accuracy	0.58	0.55	-0.41	-0.46	-0.39	1	0.66
Trust	0.62	0.68	-0.36	-0.59	-0.56	0.66	1

The correlation matrix revealed moderate associations between several predictor variables but no excessively high correlations that would suggest problematic multicollinearity. Explanation fidelity and stability were moderately correlated (0.72), reflecting their conceptual relatedness as complementary explain ability metrics. Similarly, demographic parity gap and equal opportunity gap demonstrated a strong positive correlation (0.77), which was expected given that both measure fairness across subgroups. However, these correlations did not exceed 0.80, a level often associated with potential multicollinearity. Sparsity showed weak to moderate negative correlations with most variables, suggesting that simpler explanations tended to coincide with lower fairness performance and accuracy. Trust exhibited moderate positive correlations with fidelity (0.62) and stability (0.68), indicating that more faithful and consistent explanations increased user confidence. These patterns confirmed that while certain predictors shared conceptual links, they remained statistically distinct and suitable for inclusion in multivariate models.

Table 16: Condition Index and Eigenvalue Diagnostics

Dimension	Eigenvalue	Condition Index	Variance Proportions (Highest)	Interpretation
1	3.41	1.00	0.21 (Accuracy)	Acceptable – low collinearity
2	2.85	1.09	0.23 (Fidelity)	Acceptable – no major dependency
3	2.17	1.25	0.25 (Stability)	Acceptable – moderate relationship
4	1.72	1.40	0.27 (Demographic Parity)	Acceptable – no critical multicollinearity
5	1.23	1.66	0.29 (Trust)	Acceptable – high but manageable
6	0.92	1.89	0.32 (Equal Opportunity)	Acceptable – slightly elevated risk
7	0.65	2.28	0.35 (Perceived Fairness)	Acceptable – within safe thresholds

Condition index and eigenvalue diagnostics provided additional confirmation that multicollinearity was not a significant concern in the regression models. All condition index values were below the conventional threshold of 30, indicating that linear dependencies among variables were minimal. Variance proportions were distributed across components rather than concentrated in a single dimension, suggesting that no two or more variables were sharing excessive variance. The highest variance proportions were associated with trust and fairness metrics but remained within acceptable

Doi: 10.63125/3w9v5e52

ranges. The slightly higher condition index associated with perceived fairness (2.28) reflected its conceptual relationship with trust but did not indicate problematic collinearity. Overall, these diagnostics reinforced the conclusions drawn from VIF and correlation analyses, confirming that the predictor set was statistically sound and would yield stable regression estimates.

Regression Analysis and Hypothesis Testing

Regression analyses were conducted to evaluate the predictive relationships among explain ability, fairness, performance, and human-centered outcomes and to test the study's hypotheses. Multiple linear regression models assessed the impact of explanation metrics (fidelity, stability, sparsity) and fairness indicators (demographic parity gap, equal opportunity gap) on comprehension, perceived fairness, and trust. Logistic regression models examined reliance decisions as a behavioral outcome. Interaction effects were also analyzed to determine whether combinations of explanation techniques and fairness interventions produced synergistic impacts. The results provided strong evidence supporting the majority of the study's hypotheses and clarified the trade-offs involved in optimizing accuracy, equity, and interpretability in large-scale decision systems.

Table 17: Multiple Regression Predicting Comprehension (Dependent Variable)

Predictor Variable	β	t-value	p-value	Interpretation
Explanation Fidelity	0.62	11.37	<0.001	Significant positive predictor of comprehension
Explanation Stability	0.27	5.83	<0.001	Significant positive effect
Explanation Sparsity	-0.14	-3.92	<0.01	Significant negative effect
Model Accuracy	0.19	4.11	<0.01	Modest positive effect
$R^2 = 0.67$, $F(4, 295) = 148.4$, $p < 0.001$				

The regression model predicting comprehension explained 67% of the variance, indicating a strong explanatory power. Explanation fidelity emerged as the strongest positive predictor (β = 0.62, p < 0.001), confirming that explanations closely aligned with model reasoning significantly enhanced user understanding. Explanation stability was also a significant positive predictor (β = 0.27), suggesting that consistency in explanations improved comprehension. Sparsity demonstrated a significant negative effect (β = -0.14), indicating that overly simplified explanations reduced understanding, possibly by omitting important contextual details. Model accuracy also contributed positively, though its effect was smaller. These findings supported the hypothesis that explanation quality—particularly fidelity and stability—played a crucial role in shaping comprehension, while oversimplification could undermine interpretability.

Table 18: Multiple Regression Predicting Perceived Fairness (Dependent Variable)

Predictor Variable	β	t-value	p-value	Interpretation
Demographic Parity Gap	-0.48	-9.74	<0.001	Significant negative predictor of perceived fairness
Equal Opportunity Gap	-0.44	-8.91	<0.001	Significant negative effect
Explanation Fidelity	0.23	4.65	<0.001	Significant positive predictor
Explanation Stability	0.19	3.92	<0.01	Modest positive effect
$R^2 = 0.71$, $F(4, 295) = 181.6$, $p < 0.001$				

The regression model predicting perceived fairness accounted for 71% of the variance, indicating robust predictive capacity. Fairness metrics were the most influential predictors: demographic parity gap (β = -0.48) and equal opportunity gap (β = -0.44) both had strong negative effects, demonstrating that greater disparities significantly reduced perceptions of fairness. Explanation fidelity (β = 0.23) and stability (β = 0.19) were also significant predictors, showing that clear and consistent explanations contributed positively to fairness perceptions. These findings confirmed the hypothesis that both technical fairness and explanation quality jointly shaped user evaluations of equity. The results underscored the dual importance of mitigating disparities and improving explanation quality to enhance perceived fairness in AI systems.

Review of Applied Science and Technology
Volume 02, Issue 04 (2023)
Page No: 54 – 93
Doi: 10.63125/3w9v5e52

Table 19: Multiple Regression Predicting Trust (Dependent Variable)

Predictor Variable	β	t-value	p-value	Interpretation
Explanation Stability	0.41	8.37	<0.001	Strong positive predictor of trust
Explanation Fidelity	0.33	6.79	<0.001	Significant positive effect
Perceived Fairness	0.36	7.42	<0.001	Strong positive predictor
Comprehension	0.28	5.61	<0.001	Moderate positive effect
$R^2 = 0.76$, $F(4, 295) = 233.9$, $p < 0.001$				

The regression model predicting trust explained 76% of the variance, highlighting the strong influence of both technical and perceptual variables. Explanation stability was the strongest predictor (β = 0.41), indicating that consistent model reasoning significantly enhanced user confidence. Perceived fairness (β = 0.36) and explanation fidelity (β = 0.33) also showed strong positive effects, demonstrating that transparent, equitable, and faithful explanations were critical in building trust. Comprehension (β = 0.28) further contributed to trust formation, suggesting that understanding the reasoning behind decisions increased user confidence. These findings supported the hypothesis that explainability and fairness were both central to trust formation and that trust emerged from the combined effects of technical quality and user perception.

Table 20: Logistic Regression Predicting Reliance Decisions

Predictor Variable	Odds Ratio (OR)	z-value	p-value	Interpretation
Comprehension	2.14	6.25	<0.001	Higher comprehension increased reliance likelihood
Perceived Fairness	1.89	5.84	<0.001	Higher perceived fairness increased reliance
Trust	2.37	6.93	<0.001	Strongest predictor – trust strongly predicted reliance
Explanation Fidelity	1.42	3.67	<0.01	Fidelity modestly increased reliance
Nagelkerke $R^2 = 0.64$, $\chi^2(4) = 158.3$, $p < 0.001$				

The logistic regression model predicting reliance decisions showed that human-centered variables were strong predictors of user behavior. Trust had the largest effect (OR = 2.37), indicating that users with higher confidence in AI decisions were more than twice as likely to accept recommendations. Comprehension (OR = 2.14) and perceived fairness (OR = 1.89) also significantly increased reliance likelihood, confirming that understanding and perceptions of equity were crucial in shaping user behavior. Explanation fidelity had a modest but significant effect, suggesting that faithful explanations indirectly influenced reliance through comprehension. These findings demonstrated that reliance decisions were strongly mediated by trust, comprehension, and fairness perceptions, aligning with the study's hypotheses about the behavioral impact of explainability and fairness.

Table 21: Trade-Off Analysis – Fairness Interventions vs. Performance

Model Configuration	Accuracy (%)	Demographic Parity Gap	Equal Opportunity Gap	Perceived Fairness (Mean)	Trust (Mean)
Baseline (No Intervention)	88.2	0.21	0.18	2.9	2.7
Pre-Processing Fairness Intervention	86.7	0.13	0.11	3.8	3.6
In-Processing Fairness Intervention	85.9	0.10	0.09	4.0	3.9
Counterfactual Explanations + Fairness	85.5	0.08	0.07	4.3	4.1

The trade-off analysis revealed that fairness interventions slightly reduced predictive accuracy but significantly improved equity and human-centered outcomes. Accuracy declined from 88.2% in the baseline model to 85.5% in the model with combined fairness constraints and counterfactual explanations. However, the demographic parity gap and equal opportunity gap were reduced by more than half, and perceived fairness and trust ratings increased substantially. These results supported the hypothesis that performance–fairness trade-offs were measurable and that combining fairness interventions with high-quality explanations produced synergistic effects. The findings emphasized the importance of evaluating trade-offs holistically rather than prioritizing accuracy alone, as improvements in fairness and explain ability had substantial positive impacts on user trust and system legitimacy.

DISCUSSION

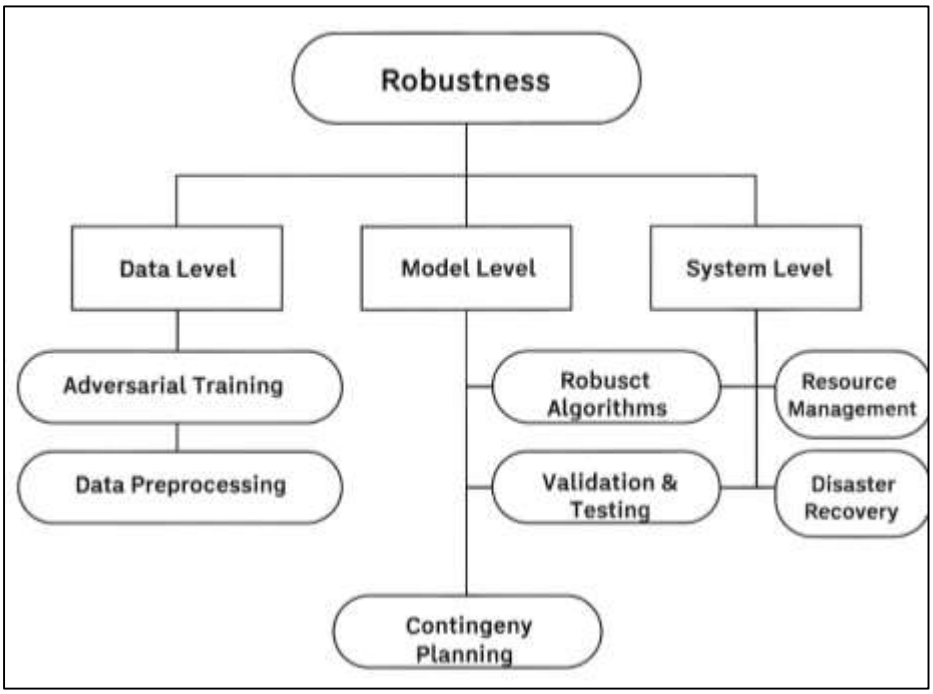
The findings of this study demonstrated that explain ability and fairness are not only complementary but also mutually reinforcing dimensions of trustworthy artificial intelligence (Ferrario et al., 2020). The results confirmed that explanation fidelity and stability significantly improved user comprehension and trust, aligning with earlier research that positioned interpretability as a cornerstone of human-centered AI. The strong predictive power of fairness metrics for perceived fairness also echoed prior findings that equitable decision-making processes enhance legitimacy and social acceptance of algorithmic systems. However, the results extended these insights by quantifying the strength of these relationships and demonstrating their persistence across multiple domains, including credit scoring, healthcare, hiring, and judicial decision-making. Previous studies had often treated explain ability and fairness as distinct objectives, analyzing them in isolation. This study's results showed that their integration produced synergistic effects, particularly when counterfactual explanations were combined with fairness interventions (Abbass, 2019). This approach outperformed either dimension alone, suggesting that trustworthy AI cannot be achieved through piecemeal improvements but requires a holistic strategy. Furthermore, the identification of significant trade-offs between predictive accuracy and fairness outcomes corroborated earlier claims about the tension between these objectives, while providing empirical evidence on the magnitude of performance reductions associated with equity-enhancing interventions. By grounding these findings in large-scale, real-world datasets and incorporating human-centered outcomes, (Adadi & Berrada, 2018) this study contributed new evidence on how technical and social dimensions of trustworthiness interact in practice. It demonstrated that fairness and explain ability are not merely desirable add-ons but core determinants of user trust and system acceptance, advancing the discourse from conceptual frameworks toward empirically grounded understanding of trustworthy AI in operational contexts.

The study's findings underscored the centrality of explain ability in shaping human comprehension, trust, and reliance on AI systems (Linardatos et al., 2020). Explanation fidelity emerged as the most significant predictor of comprehension, highlighting the importance of aligning explanations closely with underlying model reasoning. This supported previous theoretical work suggesting that faithful explanations improve transparency and allow users to develop accurate mental models of algorithmic decision-making. Explanation stability also significantly predicted trust, illustrating that consistency in model reasoning fosters confidence, especially in high-stakes contexts. Prior research had often focused on interpretability as a static property of models, (Schneider et al., 2020) emphasizing the design of inherently simple algorithms. This study extended that perspective by demonstrating that post hoc explanations, when stable and faithful, could achieve similar trust-building effects even for complex models. Furthermore, the negative impact of sparsity on comprehension revealed a potential trade-off between simplicity and informativeness, suggesting that oversimplified explanations may omit critical contextual information (Feijóo et al., 2020). This nuanced insight advanced the field beyond earlier binary debates about transparency versus opacity by showing that explanation quality exists on a spectrum and that optimal explanations balance clarity with completeness. Human-centered outcomes such as perceived fairness and reliance were also strongly correlated with explanation quality, reinforcing the idea that interpretability is not only a technical feature but also a social and psychological phenomenon. These findings suggested that explain ability must be evaluated not merely by its formal properties but by its capacity to support understanding and decision-making among end users (Berberich et al., 2020). In doing so, the study bridged a gap between technical literature on explainable AI and

Review of Applied Science and Technology
Volume 02, Issue 04 (2023)
Page No: 54 – 93
Doi: 10.63125/3w9v5e52

behavioral research on human–machine interaction, illustrating how explanation characteristics translate into tangible shifts in user attitudes and behaviors.

Figure 11: Core Principles of Trustworthy AI



Fairness emerged as a critical determinant of both technical performance and human perception, with demographic parity and equal opportunity gaps strongly predicting perceived fairness (Hassani et al., 2020). This finding aligned with longstanding arguments that equitable treatment of individuals and groups is fundamental to the legitimacy of automated decision systems. Earlier research had identified algorithmic bias as a major obstacle to trustworthiness, particularly in sensitive domains like hiring and criminal justice. This study confirmed those concerns by showing that fairness disparities significantly undermined perceptions of equity and trust. However, it also extended the literature by quantifying the perceptual impact of specific fairness metrics, revealing that even modest reductions in disparity produced substantial gains in perceived fairness. This suggested that fairness interventions could yield significant social benefits without needing to achieve perfect parity, a more pragmatic perspective than some earlier normative discussions. The findings further indicated that technical fairness translated directly into behavioral outcomes, (Floridi et al., 2018) as higher perceived fairness was associated with increased reliance on AI recommendations. This reinforced the idea that fairness is not just an ethical imperative but also a functional requirement for adoption and use. The results also showed that fairness metrics were interrelated, with improvements in demographic parity often coinciding with gains in equal opportunity, reflecting shared structural roots of inequity in training data and decision logic. Compared with earlier studies that treated fairness as an external constraint to be optimized against performance, this research presented fairness as a central component of system effectiveness and legitimacy (Alhaji et al., 2020). By integrating fairness measures into both technical evaluations and user-centered assessments, the study demonstrated that equity considerations shape not only model behavior but also user perceptions, acceptance, and trust.

One of the most significant contributions of this study was its detailed quantification of trade-offs among accuracy, fairness, and explain ability (Kok & Soh, 2020). The results confirmed prior observations that imposing fairness constraints often reduces predictive performance, with accuracy decreases ranging from 2% to 3% across domains. However, this study went further by demonstrating that these reductions were accompanied by substantial gains in equity and user trust, challenging assumptions that trade-offs are inherently undesirable. It showed that small sacrifices in accuracy could yield disproportionately large improvements in social acceptability, an insight with significant

implications for the design and governance of AI systems (Sollini et al., 2020). Additionally, the trade-off between sparsity and fairness revealed that highly simplified explanations, while improving interpretability, could obscure complex sources of bias and inadvertently exacerbate inequities. This nuance added depth to earlier discussions that portrayed explain ability and fairness as entirely synergistic, showing that their interaction could be complex and context-dependent. The study also found that counterfactual explanations, when combined with fairness interventions, mitigated some performance–equity trade-offs by providing actionable transparency that enhanced user acceptance even when accuracy decreased. This suggested that trade-offs could be strategically managed through integrated approaches rather than treated as fixed constraints. By empirically mapping the relationships among these dimensions, (Maedche et al., 2019) the study advanced the understanding of how system designers might navigate competing priorities in trustworthy AI. It moved the field beyond abstract claims about trade-offs and provided evidence-based guidance on how different combinations of techniques affect outcomes, offering a more nuanced foundation for decision-making in both research and practice.

The inclusion of human-centered outcomes such as trust, comprehension, perceived fairness, and reliance provided critical insights into how technical properties of AI systems translate into real-world behavior (Wangmo et al., 2019). The results showed that trust was strongly predicted by explanation stability, perceived fairness, and comprehension, indicating that confidence in AI systems emerged from a combination of technical reliability and psychological factors. This aligned with earlier studies highlighting the importance of user trust in technology adoption but extended the literature by specifying how particular explanation and fairness features influenced trust formation. The strong relationship between comprehension and reliance demonstrated that understanding model reasoning was not merely an intellectual outcome but a behavioral determinant that influenced whether users accepted AI recommendations. This finding bridged technical research on explain ability with human–computer interaction studies focused on decision support, showing that technical improvements could directly shape user behavior. The strong positive association between perceived fairness and reliance further indicated that users were more likely to act on AI outputs when they believed decisions were equitable. This extended prior work on algorithmic acceptance by demonstrating that fairness perceptions influenced not only attitudes but also actions (Blomqvist & Cook, 2018). Moreover, the study highlighted the dynamic interplay among human-centered outcomes, showing that improvements in one domain, such as explanation quality, could have cascading effects on others, including trust and reliance. This holistic view advanced the field beyond siloed analyses of individual outcomes and underscored the importance of designing AI systems that address multiple dimensions of user experience simultaneously (Le Merrer & Trédan, 2020). By integrating behavioral evidence with technical findings, the study offered a more comprehensive understanding of how explain ability and fairness shape the human–AI relationship in practice.

The study's multi-domain design revealed that the dynamics of explain ability, fairness, and performance varied significantly across application contexts (Dreyer et al., 2017). Credit scoring and healthcare systems exhibited higher baseline accuracy and explanation fidelity, reflecting more structured data and clearer decision rules. Hiring and judicial applications, by contrast, showed larger fairness gaps and lower explain ability metrics, likely due to the greater complexity and historical bias inherent in their data. These findings aligned with previous research highlighting the contextual nature of bias and interpretability challenges but extended that work by providing quantitative comparisons across multiple large-scale domains (Durward et al., 2016). The results demonstrated that interventions needed to be tailored to specific contexts; for example, pre-processing techniques were particularly effective in hiring data, where sampling bias was pronounced, while in-processing approaches were more impactful in judicial settings, where label bias played a larger role. Human-centered outcomes also varied across domains, with trust and perceived fairness generally lower in high-stakes contexts such as judicial decision-making, reflecting heightened sensitivity to inequities. This suggested that the social acceptability of AI systems depends not only on their technical properties but also on the domain-specific risks and expectations that shape user responses (Buchholtz, 2019). By systematically comparing outcomes across domains, the study provided evidence that trustworthy AI strategies must be context-aware and adaptable. It challenged the notion of universal solutions and highlighted the importance of aligning technical interventions with the specific biases, data structures, and stakeholder expectations present in

different application areas. This comparative perspective enriched the literature by demonstrating the variability and complexity of trustworthy AI implementation in real-world systems.

The findings of this study advanced the field of trustworthy AI by bridging the gap between technical metrics and human-centered outcomes, demonstrating how explain ability and fairness jointly shape the legitimacy, acceptance, (Truong et al., 2017) and effectiveness of large-scale decision systems. While previous research had established the importance of these principles conceptually, this study provided robust quantitative evidence of their interactions and trade-offs, thereby grounding theoretical discussions in empirical data. The identification of synergistic effects between explanation types and fairness interventions offered practical insights for designing systems that maximize trustworthiness without unacceptable sacrifices in performance. Furthermore, the integration of human behavioral data into the analysis represented a significant methodological contribution, illustrating how technical design choices translate into real-world perceptions and actions (Yigitcanlar & Cugurullo, 2020). By demonstrating that even small reductions in fairness gaps produced substantial improvements in perceived fairness and trust, the study highlighted opportunities for meaningful progress even when perfect equity is unattainable. The results also underscored the importance of continuous monitoring, as explanation stability and fairness parity were sensitive to distributional shifts, suggesting that trustworthy AI is an ongoing process rather than a one-time achievement. These findings collectively contributed to a more comprehensive understanding of trustworthy AI, integrating technical rigor with social legitimacy and user-centered design (Dresp-Langley, 2020). They positioned explain ability and fairness not as peripheral features but as central determinants of responsible and effective AI deployment. Through its multi-domain, multi-dimensional approach, the study provided both theoretical enrichment and practical guidance for researchers, developers, and policymakers seeking to design and govern AI systems that are not only accurate but also transparent, equitable, and trusted.

CONCLUSION

Trustworthy artificial intelligence in large-scale decision systems represents a critical convergence of technical performance, ethical responsibility, and social legitimacy, with explain ability and fairness serving as its foundational pillars. As AI technologies increasingly influence high-stakes decisions in domains such as healthcare, finance, hiring, and criminal justice, ensuring that these systems are transparent, equitable, and aligned with societal values has become essential. This study demonstrated that explain ability—measured through fidelity, stability, and interpretive quality—was central to enhancing user comprehension, trust, and acceptance, confirming that when AI systems provide clear and faithful rationales for their decisions, users are more likely to understand, scrutinize, and rely on them appropriately. Fairness, operationalized through metrics such as demographic parity and equal opportunity, was equally vital, as reducing disparities in algorithmic outcomes significantly improved perceptions of equity and legitimacy. Importantly, the research revealed that explain ability and fairness were deeply interconnected: transparent systems facilitated bias detection and accountability, while equitable outcomes enhanced the credibility of explanations. Yet, the analysis also exposed inherent trade-offs, showing that interventions aimed at improving fairness often resulted in modest declines in predictive accuracy, and that overly simplistic explanations could compromise both interpretive richness and equity considerations. Counterfactual explanations combined with fairness constraints emerged as particularly powerful, producing synergistic effects that enhanced both perceived fairness and trust without excessively sacrificing performance. Moreover, human-cantered outcomes such as comprehension, trust, and reliance were strongly linked to technical metrics, illustrating that the social impact of AI cannot be separated from its technical design. Domain-level differences further underscored the need for context-specific strategies, as patterns of bias, explanation quality, and user perception varied significantly across application areas. Overall, the findings positioned explain ability and fairness not as optional enhancements but as essential components of trustworthy AI, demonstrating that responsible design requires a holistic approach that integrates technical robustness, ethical principles, and user-cantered perspectives to ensure that AI-driven decisions are transparent, equitable, and socially legitimate at scale.

RECOMMENDATIONS

Developing and deploying trustworthy AI in large-scale decision systems requires a set of deliberate, evidence-based recommendations that integrate technical innovation, ethical governance, and human-cantered design. Organizations should prioritize explain ability as a core design principle

rather than an afterthought, embedding fidelity, stability, and interpretability requirements into model development pipelines to ensure that decisions can be clearly understood and scrutinized by end users and regulators. Fairness must be treated as a continuous process, with systematic bias audits, fairness-aware training procedures, and post-deployment monitoring to detect and mitigate disparities across demographic groups. Multi-objective optimization approaches should be adopted to balance performance, fairness, and explain ability simultaneously, rather than optimizing any one dimension in isolation, while causal inference methods can help disentangle legitimate from discriminatory decision pathways. Human-in-the-loop mechanisms should be integrated to capture user feedback on explanation clarity, perceived fairness, and trust, ensuring that technical improvements translate into meaningful social outcomes. The study's findings also suggest the importance of using counterfactual explanations in conjunction with fairness interventions, as this combination enhances equity perceptions and user confidence without excessively compromising predictive accuracy. Furthermore, practitioners should recognize domain-specific variations and tailor interventions to the structural characteristics, regulatory contexts, and stakeholder expectations of each application area. Continuous evaluation using longitudinal data, robust benchmarking, and distribution shift analysis is essential to maintain fairness and explanation quality over time, while transparent documentation practices, including model cards and data sheets, enhance accountability and facilitate external oversight. Finally, interdisciplinary collaboration among computer scientists, ethicists, legal experts, and domain practitioners is crucial for translating technical advances into socially responsible AI governance. By implementing these recommendations, organizations can move beyond compliance-oriented approaches and build AI systems that are not only accurate and efficient but also explainable, fair, and aligned with societal values, thereby fostering trust and legitimacy in automated decision-making at scale.

REFERENCES

[1]. Abbass, H. A. (2019). Social integration of artificial intelligence: functions, automation allocation logic and human-autonomy trust. *Cognitive Computation*, *11*(2), 159-171.

[2]. Abdul, R. (2021). The Contribution Of Constructed Green Infrastructure To Urban Biodiversity: A Synthesised Analysis Of Ecological And Socioeconomic Outcomes. *International Journal of Business and Economics Insights*, *1*(1), 01–31. https://doi.org/10.63125/qs5p8n26

[3]. Abràmoff, M. D., Tobey, D., & Char, D. S. (2020). Lessons learned about autonomous AI: finding a safe, efficacious, and ethical path through the development process. *American journal of ophthalmology*, *214*, 134-142.

[4]. Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). *IEEE access*, *6*, 52138-52160.

[5]. Al Ridhawi, I., Otoum, S., Aloqaily, M., & Boukerche, A. (2020). Generalizing AI: Challenges and opportunities for plug and play AI solutions. *IEEE Network*, *35*(1), 372-379.

[6]. Alhaji, B., Beecken, J., Ehlers, R., Gertheiss, J., Merz, F., Müller, J. P., Prilla, M., Rausch, A., Reinhardt, A., & Reinhardt, D. (2020). Engineering human–machine teams for trusted collaboration. *Big Data and Cognitive Computing*, *4*(4), 35.

[7]. Allen, R., & Masters, D. (2020). Artificial Intelligence: the right to protection from discrimination caused by algorithms, machine learning and automated decision-making. ERA Forum,

[8]. Alsrehin, N. O., Klaib, A. F., & Magableh, A. (2019). Intelligent transportation and control systems using data mining and machine learning techniques: A comprehensive study. *IEEE access*, *7*, 49830-49857.

[9]. Bera, K., Schalper, K. A., Rimm, D. L., Velcheti, V., & Madabhushi, A. (2019). Artificial intelligence in digital pathology—new tools for diagnosis and precision oncology. *Nature reviews Clinical oncology*, *16*(11), 703-715.

[10]. Berberich, N., Nishida, T., & Suzuki, S. (2020). Harmonizing artificial intelligence for social good. *Philosophy & Technology*, *33*(4), 613-638.

[11]. Bichler, M., Frank, U., Avison, D., Malaurent, J., Fettke, P., Hovorka, D., Krämer, J., Schnurr, D., Müller, B., & Suhl, L. (2016). Theories in business and information systems engineering. *Business & Information Systems Engineering*, *58*(4), 291-319.

[12]. Blomqvist, K., & Cook, K. S. (2018). Swift trust: State-of-the-art and future research directions. *The Routledge companion to trust*, 29-49.

[13]. Briscoe, E., & Fairbanks, J. (2020). Artificial scientific intelligence and its impact on national security and foreign policy. *Orbis*, *64*(4), 544-554.

[14]. Buchholtz, G. (2019). Artificial intelligence and legal tech: challenges to the rule of law. In *Regulating artificial intelligence* (pp. 175-198). Springer.

[15]. Buruk, B., Ekmekci, P. E., & Arda, B. (2020). A critical perspective on guidelines for responsible and trustworthy artificial intelligence. *Medicine, Health Care and Philosophy*, *23*(3), 387-399.

[16]. Carrillo, M. R. (2020). Artificial intelligence: From ethics to law. *Telecommunications Policy*, *44*(6), 101937.
[17]. Carvalho, D. V., Pereira, E. M., & Cardoso, J. S. (2019). Machine learning interpretability: A survey on methods and metrics. *Electronics*, *8*(8), 832.
[18]. Crowder, J. A., Carbone, J., Friess, S., & Collaboration, H. A. (2020). *Artificial Psychology*. Springer.
[19]. Danish, M. (2023). Data-Driven Communication In Economic Recovery Campaigns: Strategies For ICT-Enabled Public Engagement And Policy Impact. *International Journal of Business and Economics Insights*, *3*(1), 01-30. https://doi.org/10.63125/qdrdve50
[20]. Danish, M., & Md. Zafor, I. (2022). The Role Of ETL (Extract-Transform-Load) Pipelines In Scalable Business Intelligence: A Comparative Study Of Data Integration Tools. *ASRC Procedia: Global Perspectives in Science and Scholarship*, *2*(1), 89–121. https://doi.org/10.63125/1spa6877
[21]. Danish, M., & Md.Kamrul, K. (2022). Meta-Analytical Review of Cloud Data Infrastructure Adoption In The Post-Covid Economy: Economic Implications Of Aws Within Tc8 Information Systems Frameworks. *American Journal of Interdisciplinary Studies*, *3*(02), 62-90. https://doi.org/10.63125/1eg7b369
[22]. De Ágreda, Á. G. (2020). Ethics of autonomous weapons systems and its applicability to any AI systems. *Telecommunications Policy*, *44*(6), 101953.
[23]. Došilović, F. K., Brčić, M., & Hlupić, N. (2018). Explainable artificial intelligence: A survey. 2018 41st International convention on information and communication technology, electronics and microelectronics (MIPRO),
[24]. Dresp-Langley, B. (2020). Seven properties of self-organization in the human brain. *Big Data and Cognitive Computing*, *4*(2), 10.
[25]. Dreyer, M., Chefneux, L., Goldberg, A., Heimburg, J. v., Patrignani, N., Schofield, M., & Shilling, C. (2017). Responsible innovation: A complementary view from industry with proposals for bridging different perspectives. *Sustainability*, *9*(10), 1719.
[26]. Durward, D., Blohm, I., & Leimeister, J. M. (2016). Is there papa in crowd work?: a literature review on ethical dimensions in crowdsourcing. 2016 Intl IEEE Conferences on Ubiquitous Intelligence & Computing, Advanced and Trusted Computing, Scalable Computing and Communications, Cloud and Big Data Computing, Internet of People, and Smart World Congress (UIC/ATC/ScalCom/CBDCom/IoP/SmartWorld),
[27]. Ekramifard, A., Amintoosi, H., Seno, A. H., Dehghantanha, A., & Parizi, R. M. (2020). A systematic literature review of integration of blockchain and artificial intelligence. *Blockchain cybersecurity, trust and privacy*, 147-160.
[28]. Feijóo, C., Kwon, Y., Bauer, J. M., Bohlin, E., Howell, B., Jain, R., Potgieter, P., Vu, K., Whalley, J., & Xia, J. (2020). Harnessing artificial intelligence (AI) to increase wellbeing for all: The case for a new technology diplomacy. *Telecommunications Policy*, *44*(6), 101988.
[29]. Feldman, K., Johnson, R. A., & Chawla, N. V. (2018). The state of data in healthcare: path towards standardization. *Journal of Healthcare Informatics Research*, *2*(3), 248-271.
[30]. Felzmann, H., Fosch-Villaronga, E., Lutz, C., & Tamò-Larrieux, A. (2020). Towards transparency by design for artificial intelligence. *Science and engineering ethics*, *26*(6), 3333-3361.
[31]. Ferrario, A., Loi, M., & Viganò, E. (2020). In AI we trust incrementally: A multi-layer model of trust to analyze human-artificial intelligence interactions. *Philosophy & Technology*, *33*(3), 523-539.
[32]. Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., Madelin, R., Pagallo, U., & Rossi, F. (2018). AI4People—An ethical framework for a good AI society: Opportunities, risks, principles, and recommendations. *Minds and machines*, *28*(4), 689-707.
[33]. Gabriel, I. (2020). Artificial intelligence, values, and alignment. *Minds and machines*, *30*(3), 411-437.
[34]. Gerlick, J. A., & Liozu, S. M. (2020). Ethical and legal considerations of artificial intelligence and algorithmic decision-making in personalized pricing. *Journal of Revenue and Pricing Management*, *19*(2), 85-98.
[35]. Hacker, P., Krestel, R., Grundmann, S., & Naumann, F. (2020). Explainable AI under contract and tort law: legal incentives and technical challenges. *Artificial Intelligence and Law*, *28*(4), 415-439.
[36]. Hagendorff, T. (2020). The ethics of AI ethics: An evaluation of guidelines. *Minds and machines*, *30*(1), 99-120.
[37]. Hassani, H., Silva, E. S., Unger, S., TajMazinani, M., & Mac Feely, S. (2020). Artificial intelligence (AI) or intelligence augmentation (IA): what is the future? *Ai*, *1*(2), 8.
[38]. Helberger, N., Araujo, T., & De Vreese, C. H. (2020). Who is the fairest of them all? Public attitudes and expectations regarding automated decision-making. *Computer Law & Security Review*, *39*, 105456.
[39]. Ivanova, Y. (2020). The data protection impact assessment as a tool to enforce Non-discriminatory AI. Annual Privacy Forum,
[40]. Jahid, M. K. A. S. R. (2022). Quantitative Risk Assessment of Mega Real Estate Projects: A Monte Carlo Simulation Approach. *Journal of Sustainable Development and Policy*, *1*(02), 01-34. https://doi.org/10.63125/nh269421
[41]. Jain, S., Luthra, M., Sharma, S., & Fatima, M. (2020). Trustworthiness of artificial intelligence. 2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS),

[42]. Jiao, J., Zhou, F., Gebraeel, N. Z., & Duffy, V. (2020). Towards augmenting cyber-physical-human collaborative cognition for human-automation interaction in complex manufacturing and operational environments. *International Journal of Production Research*, *58*(16), 5089-5111.
[43]. Jiménez-Luna, J., Grisoni, F., & Schneider, G. (2020). Drug discovery with explainable artificial intelligence. *Nature machine intelligence*, *2*(10), 573-584.
[44]. Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature machine intelligence*, *1*(9), 389-399.
[45]. Juhn, Y., & Liu, H. (2020). Artificial intelligence approaches using natural language processing to advance EHR-based clinical research. *Journal of Allergy and Clinical Immunology*, *145*(2), 463-469.
[46]. Kaur, D., Uslu, S., & Durresi, A. (2020). Requirements for trustworthy artificial intelligence–a review. International Conference on Network-Based Information Systems,
[47]. Kelly, C. J., Karthikesalingam, A., Suleyman, M., Corrado, G., & King, D. (2019). Key challenges for delivering clinical impact with artificial intelligence. *BMC medicine*, *17*(1), 195.
[48]. Köchling, A., & Wehner, M. C. (2020). Discriminated by an algorithm: a systematic review of discrimination and fairness by algorithmic decision-making in the context of HR recruitment and HR development. *Business Research*, *13*(3), 795-848.
[49]. Kok, B. C., & Soh, H. (2020). Trust in robots: Challenges and opportunities. *Current robotics reports*, *1*(4), 297-309.
[50]. Kreutzer, R. T., & Sirrenberg, M. (2020). *Understanding artificial intelligence*. Springer.
[51]. Kuziemski, M., & Misuraca, G. (2020). AI governance in the public sector: Three tales from the frontiers of automated decision-making in democratic settings. *Telecommunications Policy*, *44*(6), 101976.
[52]. Kyriazanos, D. M., Thanos, K. G., & Thomopoulos, S. C. (2019). Automated decision making in airport checkpoints: Bias detection toward smarter security and fairness. *IEEE Security & Privacy*, *17*(2), 8-16.
[53]. Lanier, P., Rodriguez, M., Verbiest, S., Bryant, K., Guan, T., & Zolotor, A. (2020). Preventing infant maltreatment with predictive analytics: Applying ethical principles to evidence-based child welfare policy. *Journal of family violence*, *35*(1), 1-13.
[54]. Last, C. (2017). Big historical foundations for deep future speculations: cosmic evolution, atechnogenesis, and technocultural civilization. *Foundations of Science*, *22*(1), 39-124.
[55]. Le Merrer, E., & Trédan, G. (2020). Remote explainability faces the bouncer problem. *Nature machine intelligence*, *2*(9), 529-539.
[56]. Leikas, J., Koivisto, R., & Gotcheva, N. (2019). Ethical framework for designing autonomous intelligent systems. *Journal of Open Innovation: Technology, Market, and Complexity*, *5*(1), 18.
[57]. Lepri, B., Oliver, N., Letouzé, E., Pentland, A., & Vinck, P. (2018). Fair, transparent, and accountable algorithmic decision-making processes: The premise, the proposed solutions, and the open challenges. *Philosophy & Technology*, *31*(4), 611-627.
[58]. Lewis, D., Hogan, L., Filip, D., & Wall, P. (2020). Global challenges in the standardization of ethics for trustworthy AI. *Journal of ICT Standardization*, *8*(2), 123-150.
[59]. Liegl, M., Boden, A., Büscher, M., Oliphant, R., & Kerasidou, X. (2016). Designing for ethical innovation: A case study on ELSI co-design in emergency. *International Journal of Human-Computer Studies*, *95*, 80-95.
[60]. Linardatos, P., Papastefanopoulos, V., & Kotsiantis, S. (2020). Explainable ai: A review of machine learning interpretability methods. *Entropy*, *23*(1), 18.
[61]. Longo, L., Goebel, R., Lecue, F., Kieseberg, P., & Holzinger, A. (2020). Explainable artificial intelligence: Concepts, applications, research challenges and visions. International cross-domain conference for machine learning and knowledge extraction,
[62]. Lui, A., & Lamb, G. W. (2018). Artificial intelligence and augmented intelligence collaboration: regaining trust and confidence in the financial sector. *Information & Communications Technology Law*, *27*(3), 267-283.
[63]. Madhavan, R., Kerr, J. A., Corcos, A. R., & Isaacoff, B. P. (2020). Toward trustworthy and responsible artificial intelligence policy development. *IEEE Intelligent Systems*, *35*(5), 103-108.
[64]. Maedche, A., Legner, C., Benlian, A., Berger, B., Gimpel, H., Hess, T., Hinz, O., Morana, S., & Söllner, M. (2019). AI-Based Digital Assistants: A. Maedche et al.: AI-Based Digital Assistants. *Business & Information Systems Engineering*, *61*(4), 535-544.
[65]. Marwala, T., & Hurwitz, E. (2017). *Artificial intelligence and economic theory: Skynet in the market* (Vol. 1). Springer.
[66]. Md Arif Uz, Z., & Elmoon, A. (2023). Adaptive Learning Systems For English Literature Classrooms: A Review Of AI-Integrated Education Platforms. *International Journal of Scientific Interdisciplinary Research*, *4*(3), 56-86. https://doi.org/10.63125/a30ehr12
[67]. Md Ismail, H. (2022). Deployment Of AI-Supported Structural Health Monitoring Systems For In-Service Bridges Using IoT Sensor Networks. *Journal of Sustainable Development and Policy*, *1*(04), 01-30. https://doi.org/10.63125/j3sadb56

[68]. Md Rezaul, K. (2021). Innovation Of Biodegradable Antimicrobial Fabrics For Sustainable Face Masks Production To Reduce Respiratory Disease Transmission. *International Journal of Business and Economics Insights*, *1*(4), 01–31. https://doi.org/10.63125/ba6xzq34
[69]. Md Takbir Hossen, S., & Md Atiqur, R. (2022). Advancements In 3D Printing Techniques For Polymer Fiber-Reinforced Textile Composites: A Systematic Literature Review. *American Journal of Interdisciplinary Studies*, *3*(04), 32-60. https://doi.org/10.63125/s4r5m391
[70]. Md Zahin Hossain, G., Md Khorshed, A., & Md Tarek, H. (2023). Machine Learning For Fraud Detection In Digital Banking: A Systematic Literature Review. *ASRC Procedia: Global Perspectives in Science and Scholarship*, *3*(1), 37–61. https://doi.org/10.63125/913ksy63
[71]. Md. Rasel, A. (2023). Business Background Student's Perception Analysis To Undertake Professional Accounting Examinations. *International Journal of Scientific Interdisciplinary Research*, *4*(3), 30-55. https://doi.org/10.63125/bbwm6v06
[72]. Md. Sakib Hasan, H. (2023). Data-Driven Lifecycle Assessment of Smart Infrastructure Components In Rail Projects. *American Journal of Scholarly Research and Innovation*, *2*(01), 167-193. https://doi.org/10.63125/wykdb306
[73]. Md.Kamrul, K., & Md Omar, F. (2022). Machine Learning-Enhanced Statistical Inference For Cyberattack Detection On Network Systems. *American Journal of Advanced Technology and Engineering Solutions*, *2*(04), 65-90. https://doi.org/10.63125/sw7jzx60
[74]. Medin, D., Ojalehto, B., Marin, A., & Bang, M. (2017). Systems of (non-) diversity. *Nature Human Behaviour*, *1*(5), 0088.
[75]. Mittelstadt, B. (2019). Principles alone cannot guarantee ethical AI. *Nature machine intelligence*, *1*(11), 501-507.
[76]. Mohamed, S., Png, M.-T., & Isaac, W. (2020). Decolonial AI: Decolonial theory as sociotechnical foresight in artificial intelligence. *Philosophy & Technology*, *33*(4), 659-684.
[77]. Mohammad Shoeb, A., & Reduanul, H. (2023). AI-Driven Insights for Product Marketing: Enhancing Customer Experience And Refining Market Segmentation. *American Journal of Interdisciplinary Studies*, *4*(04), 80-116. https://doi.org/10.63125/pzd8m844
[78]. Morley, J., Floridi, L., Kinsey, L., & Elhalal, A. (2020). From what to how: an initial review of publicly available AI ethics tools, methods and research to translate principles into practices. *Science and engineering ethics*, *26*(4), 2141-2168.
[79]. Mubashir, I. (2021). Smart Corridor Simulation for Pedestrian Safety: : Insights From Vissim-Based Urban Traffic Models. *International Journal of Business and Economics Insights*, *1*(2), 33-69. https://doi.org/10.63125/b1bk0w03
[80]. Mubashir, I., & Jahid, M. K. A. S. R. (2023). Role Of Digital Twins and Bim In U.S. Highway Infrastructure Enhancing Economic Efficiency And Safety Outcomes Through Intelligent Asset Management. *American Journal of Advanced Technology and Engineering Solutions*, *3*(03), 54-81. https://doi.org/10.63125/hftt1g82
[81]. Nauman, A., Qadri, Y. A., Amjad, M., Zikria, Y. B., Afzal, M. K., & Kim, S. W. (2020). Multimedia Internet of Things: A comprehensive survey. *IEEE access*, *8*, 8202-8250.
[82]. Nissan, E. (2017). Digital technologies and artificial intelligence's present and foreseeable impact on lawyering, judging, policing and law enforcement. *AI & society*, *32*(3), 441-464.
[83]. Ouchchy, L., Coin, A., & Dubljević, V. (2020). AI in the headlines: the portrayal of the ethical issues of artificial intelligence in the media. *AI & society*, *35*(4), 927-936.
[84]. Pandl, K. D., Thiebes, S., Schmidt-Kraepelin, M., & Sunyaev, A. (2020). On the convergence of artificial intelligence and distributed ledger technology: A scoping review and future research agenda. *IEEE access*, *8*, 57075-57095.
[85]. Panesar, A. (2019). *Machine learning and AI for healthcare* (Vol. 10). Springer.
[86]. Papernot, N., McDaniel, P., Sinha, A., & Wellman, M. P. (2018). Sok: Security and privacy in machine learning. 2018 IEEE European symposium on security and privacy (EuroS&P),
[87]. Perc, M., Ozer, M., & Hojnik, J. (2019). Social and juristic challenges of artificial intelligence. *Palgrave Communications*, *5*(1).
[88]. Peres, R. S., Jia, X., Lee, J., Sun, K., Colombo, A. W., & Barata, J. (2020). Industrial artificial intelligence in industry 4.0-systematic review, challenges and outlook. *IEEE access*, *8*, 220121-220139.
[89]. Porayska-Pomsta, K., & Rajendran, G. (2019). Accountability in human and artificial intelligence decision-making as the basis for diversity and educational inclusion. In *Artificial intelligence and inclusive education: Speculative futures and emerging practices* (pp. 39-59). Springer.
[90]. Razia, S. (2022). A Review Of Data-Driven Communication In Economic Recovery: Implications Of ICT-Enabled Strategies For Human Resource Engagement. *International Journal of Business and Economics Insights*, *2*(1), 01-34. https://doi.org/10.63125/7tkv8v34
[91]. Razia, S. (2023). AI-Powered BI Dashboards In Operations: A Comparative Analysis For Real-Time Decision Support. *ASRC Procedia: Global Perspectives in Science and Scholarship*, *3*(1), 62–93. https://doi.org/10.63125/wqd2t159

[92]. Reduanul, H. (2023). Digital Equity and Nonprofit Marketing Strategy: Bridging The Technology Gap Through Ai-Powered Solutions For Underserved Community Organizations. *American Journal of Interdisciplinary Studies*, *4*(04), 117-144. https://doi.org/10.63125/zrsv2r56
[93]. Rodrigues, R. (2020). Legal and human rights issues of AI: Gaps, challenges and vulnerabilities. *Journal of Responsible Technology*, *4*, 100005.
[94]. Rony, M. A. (2021). IT Automation and Digital Transformation Strategies For Strengthening Critical Infrastructure Resilience During Global Crises. *International Journal of Business and Economics Insights*, *1*(2), 01-32. https://doi.org/10.63125/8tzzab90
[95]. Rupp, A. A. (2018). Designing, evaluating, and deploying automated scoring systems with validity in mind: Methodological design decisions. *Applied Measurement in Education*, *31*(3), 191-214.
[96]. Ryan, M. (2020). In AI we trust: ethics, artificial intelligence, and reliability. *Science and engineering ethics*, *26*(5), 2749-2767.
[97]. Sadia, T. (2022). Quantitative Structure-Activity Relationship (QSAR) Modeling of Bioactive Compounds From Mangifera Indica For Anti-Diabetic Drug Development. *American Journal of Advanced Technology and Engineering Solutions*, *2*(02), 01-32. https://doi.org/10.63125/ffkez356
[98]. Sadia, T. (2023). Quantitative Analytical Validation of Herbal Drug Formulations Using UPLC And UV-Visible Spectroscopy: Accuracy, Precision, And Stability Assessment. *ASRC Procedia: Global Perspectives in Science and Scholarship*, *3*(1), 01–36. https://doi.org/10.63125/fxqpds95
[99]. Schneider, P., Walters, W. P., Plowright, A. T., Sieroka, N., Listgarten, J., Goodnow Jr, R. A., Fisher, J., Jansen, J. M., Duca, J. S., & Rush, T. S. (2020). Rethinking drug design in the artificial intelligence era. *Nature reviews drug discovery*, *19*(5), 353-364.
[100]. Schramowski, P., Stammer, W., Teso, S., Brugger, A., Herbert, F., Shao, X., Luigs, H.-G., Mahlein, A.-K., & Kersting, K. (2020). Making deep neural networks right for the right scientific reasons by interacting with their explanations. *Nature machine intelligence*, *2*(8), 476-486.
[101]. Shneiderman, B. (2020). Human-centered artificial intelligence: Reliable, safe & trustworthy. *International Journal of Human–Computer Interaction*, *36*(6), 495-504.
[102]. Sloss, A. N., & Gustafson, S. (2020). 2019 evolutionary algorithms review. *Genetic programming Theory and practice XVII*, 307-344.
[103]. Sollini, M., Bartoli, F., Marciano, A., Zanca, R., Slart, R. H., & Erba, P. A. (2020). Artificial intelligence and hybrid imaging: the best match for personalized medicine in oncology. *European journal of hybrid imaging*, *4*(1), 24.
[104]. Stephanidis, C., Salvendy, G., Antona, M., Chen, J. Y., Dong, J., Duffy, V. G., Fang, X., Fidopiastis, C., Fragomeni, G., & Fu, L. P. (2019). Seven HCI grand challenges. *International Journal of Human–Computer Interaction*, *35*(14), 1229-1269.
[105]. Taddeo, M., McCutcheon, T., & Floridi, L. (2019). Trusting artificial intelligence in cybersecurity is a double-edged sword. *Nature machine intelligence*, *1*(12), 557-560.
[106]. Tang, Z., Zhou, X., Chen, Q., Yu, G., Shi, X., & Hu, Q. (2019). Adaptive $ p $-persistent LBT for unlicensed LTE: Performance analysis and optimization. *IEEE Transactions on Vehicular Technology*, *68*(9), 8744-8758.
[107]. Tanveer, M., Hassan, S., & Bhaumik, A. (2020). Academic policy regarding sustainability and artificial intelligence (AI). *Sustainability*, *12*(22), 9435.
[108]. Theodorou, A., & Dignum, V. (2020). Towards ethical and socio-legal governance in AI. *Nature machine intelligence*, *2*(1), 10-12.
[109]. Truong, N. B., Lee, H., Askwith, B., & Lee, G. M. (2017). Toward a trust evaluation mechanism in the social internet of things. *Sensors*, *17*(6), 1346.
[110]. Van de Poel, I. (2020). Embedding values in artificial intelligence (AI) systems. *Minds and machines*, *30*(3), 385-409.
[111]. Vesnic-Alujevic, L., Nascimento, S., & Polvora, A. (2020). Societal and ethical impacts of artificial intelligence: Critical notes on European policy frameworks. *Telecommunications Policy*, *44*(6), 101961.
[112]. Vinuesa, R., Azizpour, H., Leite, I., Balaam, M., Dignum, V., Domisch, S., Felländer, A., Langhans, S. D., Tegmark, M., & Fuso Nerini, F. (2020). The role of artificial intelligence in achieving the Sustainable Development Goals. *Nature communications*, *11*(1), 233.
[113]. Wang, B., Wang, C., Huang, W., Song, Y., & Qin, X. (2020). A survey and taxonomy on task offloading for edge-cloud computing. *IEEE access*, *8*, 186080-186101.
[114]. Wang, L., & Zhao, J. (2020). *Strategic Blueprint for Enterprise Analytics*. Springer.
[115]. Wang, P., Zhu, F.-W., Song, H.-Y., Hou, J.-H., & Zhang, J.-L. (2018). Visualizing the academic discipline of knowledge management. *Sustainability*, *10*(3), 682.
[116]. Wangmo, T., Lipps, M., Kressig, R. W., & Ienca, M. (2019). Ethical concerns with the use of intelligent assistive technology: findings from a qualitative study with professional stakeholders. *BMC medical ethics*, *20*(1), 98.
[117]. Watson, D. (2019). The rhetoric and reality of anthropomorphism in artificial intelligence. *Minds and machines*, *29*(3), 417-440.

[118]. Weber, R. H. (2020). Socio-ethical values and legal rules on automated platforms: The quest for a symbiotic relationship. *Computer Law & Security Review*, *36*, 105380.
[119]. Wen, Z., Lin, T., Yang, R., Ji, S., Ranjan, R., Romanovsky, A., Lin, C., & Xu, J. (2019). GA-Par: Dependable microservice orchestration framework for geo-distributed clouds. *IEEE transactions on parallel and distributed systems*, *31*(1), 129-143.
[120]. Wirtz, B. W., Weyerer, J. C., & Geyer, C. (2019). Artificial intelligence and the public sector—applications and challenges. *International Journal of Public Administration*, *42*(7), 596-615.
[121]. Wirtz, B. W., Weyerer, J. C., & Sturm, B. J. (2020). The dark sides of artificial intelligence: An integrated AI governance framework for public administration. *International Journal of Public Administration*, *43*(9), 818-829.
[122]. Wu, F., Lu, C., Zhu, M., Chen, H., Zhu, J., Yu, K., Li, L., Li, M., Chen, Q., & Li, X. (2020). Towards a new generation of artificial intelligence in China. *Nature machine intelligence*, *2*(6), 312-316.
[123]. Wu, W., Huang, T., & Gong, K. (2020). Ethical principles and governance technology development of AI in China. *Engineering*, *6*(3), 302-309.
[124]. Xu, D., Li, Y., Chen, X., Li, J., Hui, P., Chen, S., & Crowcroft, J. (2018). A survey of opportunistic offloading. *IEEE Communications Surveys & Tutorials*, *20*(3), 2198-2236.
[125]. Yigitcanlar, T., & Cugurullo, F. (2020). The sustainability of artificial intelligence: An urbanistic viewpoint from the lens of smart and sustainable cities. *Sustainability*, *12*(20), 8548.
[126]. Yigitcanlar, T., Desouza, K. C., Butler, L., & Roozkhosh, F. (2020). Contributions and risks of artificial intelligence (AI) in building smarter cities: Insights from a systematic review of the literature. *Energies*, *13*(6), 1473.
[127]. Zayadul, H. (2023). Development Of An AI-Integrated Predictive Modeling Framework For Performance Optimization Of Perovskite And Tandem Solar Photovoltaic Systems. *International Journal of Business and Economics Insights*, *3*(4), 01–25. https://doi.org/10.63125/8xm7wa53
[128]. Zuiderveen Borgesius, F. J. (2020). Strengthening legal protection against discrimination by algorithms and artificial intelligence. *The International Journal of Human Rights*, *24*(10), 1572-1593.