

Article

Investing in AI Interpretability, Control, and Robustness

Maikel Leon 

Department of Business Technology, Miami Herbert Business School, University of Miami,
Coral Gables, FL 33146, USA; mleon@miami.edu

Abstract

Artificial intelligence (AI) powers breakthroughs in language processing, computer vision, and scientific discovery; yet, the increasing complexity of frontier models makes their reasoning opaque. This opacity undermines public trust, complicates deployment in safety-critical settings, and frustrates compliance with emerging regulations. In response to initiatives such as the White House AI Action Plan, we synthesize the scientific foundations and policy landscape for interpretability, control, and robustness. We clarify key concepts and survey intrinsically interpretable and post-hoc explanation techniques, discuss human-centered evaluation and governance, and analyze how adversarial threats and distributional shifts motivate robustness research. An empirical case study compares logistic regression, random forests, and gradient boosting on a synthetic dataset with a binary-sensitive attribute using accuracy, F_1 score, and group-fairness metrics, and illustrates trade-offs between performance and fairness. We integrate ethical and policy perspectives, including recommendations from America's AI Action Plan and recent civil rights frameworks, and conclude with guidance for researchers, practitioners, and policymakers on advancing trustworthy AI.

Keywords: artificial intelligence; interpretability; explainable AI; robustness; fairness; policy

1. Introduction

Artificial intelligence (AI) exhibits the potential to revolutionize numerous domains spanning industrial production, scientific research, and everyday human activities. State-of-the-art foundation models, such as large language models and multimodal architectures, now autonomously generate coherent text, classify complex images, and accelerate drug and material discovery. These capabilities, however, result from highly non-linear and over-parameterized networks, the internal representations of which are challenging to characterize analytically. As noted in the White House AI Action Plan, the opacity of frontier models complicates their deployment in defense, national security, or other high-stakes contexts where reliability and predictability are paramount [1]. Empirical surveys and systematic reviews further indicate that stakeholders consistently request interpretable systems, yet they often prioritize predictive accuracy when lives or livelihoods are at stake [2,3]. Concurrently, research has exposed susceptibilities to adversarial perturbations and distribution shifts that jeopardize both safety and fairness [4]. Collectively, these insights have galvanized an international research agenda on explainability, controllability, and robustness in AI.

Our investigation is motivated by a tension between rapid innovation and mounting calls for transparency, fairness, and accountability. As AI systems increasingly inform credit lending, hiring, and medical diagnoses, stakeholders demand explanations that clarify how



Academic Editor: Alexander E.I. Brownlee

Received: 7 January 2026

Revised: 27 January 2026

Accepted: 6 February 2026

Published: 9 February 2026

Copyright: © 2026 by the author.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

and for whom models produce outcomes. Existing surveys tend to examine interpretability, control, and robustness separately, and policy analyses seldom integrate scientific advances into national strategies. We therefore strive to synthesize these domains, rigorously articulating conceptual definitions and technical methods while situating them within ethical and regulatory frameworks. In particular, we derive mathematical expressions for prominent post-hoc explanation techniques, such as LIME, SHAP, and integrated gradients, and evaluate their theoretical foundations. We then present a reproducible empirical case study that quantifies trade-offs among interpretability, predictive performance, and group fairness metrics, including demographic parity and equalized odds. Finally, we conduct a comparative policy analysis contrasting America's AI Action Plan and the AI Bill of Rights, elucidating common principles and normative divergences. Our holistic approach aims to guide researchers, practitioners, and policymakers in prioritizing investments and evaluating AI systems.

This article examines the three interrelated priorities: interpretability, control, and robustness (from both scientific and policy perspectives). The goal is not merely to catalog algorithms, but to integrate technical advances with normative considerations, such as fairness, accountability, and transparency [5]. Throughout, we emphasize that interpretability is not a monolithic property, but rather a continuum influenced by the system's purpose, stakeholder needs, and socio-cultural context. We also underscore that enhancing interpretability must not come at the expense of predictive performance or robustness and that investments in control and safety are essential to realize the benefits of AI responsibly. In addition to this conceptual survey, we provide an empirical case study that compares interpretable and black-box models on a synthetic dataset with a binary sensitive attribute. The case study illustrates transparency, fairness, and interpretability by analyzing differences in demographic parity and equalized odds across logistic regression, random forests, and gradient boosting models.

The rest of the paper is organized as follows. After the introduction, Section 2 describes the review methodology, including the search strategy and selection criteria used to compile the literature. Section 3 introduces conceptual foundations and clarifies terminology. Section 4 surveys technical approaches to interpretability and derives formulas for key post hoc explanation methods. Section 5 presents an empirical case study quantifying trade-offs among interpretability, performance, and fairness. Section 6 discusses mechanisms for human oversight and governance, while Section 7 examines robustness and safety. Section 8 outlines the policy landscape, contrasting America's AI Action Plan with the AI Bill of Rights, and Section 9 describes open challenges and research directions. Finally, Section 10 synthesizes the lessons and offers guidance.

2. Review Methodology

To ensure comprehensive and reproducible coverage of the rapidly evolving literature on AI interpretability, control, robustness, fairness, and governance, we conducted a structured review following scoping review guidelines. We searched multiple databases: *Web of Science*, *IEEE Xplore*, *ACM Digital Library*, *PubMed*, and *arXiv*, for publications between 2018 and 2025 using combinations of keywords such as “explainable artificial intelligence”, “interpretability”, “fairness”, “robustness”, “adversarial training”, “mechanistic interpretability”, and “AI governance”. The search was complemented by backward and forward citation chaining on seminal papers. Inclusion criteria required that papers discuss technical methods, evaluation frameworks, or policy guidelines related to interpretability, control, robustness, fairness, or governance. We excluded works that solely provided opinion pieces or lacked methodological detail. The first author screened titles and abstracts, and full texts were reviewed when relevance was uncertain. The complete

procedure for article retrieval and screening is outlined in Algorithm 1. The final corpus comprised 125 peer-reviewed articles, seven reports and standards, and two preprints (to capture very recent developments); full bibliographic details appear in the reference list. This methodology enables transparency about the scope and limitations of our review and reduces selection bias by drawing from diverse sources.

Algorithm 1 Systematic review procedure

```

1: Input: set of databases  $D$ , keyword list  $K$ , time interval  $T$ , inclusion criteria  $C_{inc}$ ,
   exclusion criteria  $C_{exc}$ 
2: Output: corpus  $\mathcal{L}$ 
3: Initialize  $\mathcal{L} \leftarrow \emptyset$  and temporary list  $\mathcal{L}_{temp} \leftarrow \emptyset$ 
4: for each database  $d \in D$  do
5:   Query  $d$  with keywords  $K$  and interval  $T$  to obtain candidate set  $S_d$ 
6:   Append  $S_d$  to  $\mathcal{L}_{temp}$ 
7: end for
8: Remove duplicate articles from  $\mathcal{L}_{temp}$  to get  $\mathcal{L}_{unique}$ 
9: for each article  $a \in \mathcal{L}_{unique}$  do
10:  Screen the title and abstract of  $a$ 
11:  if  $a$  satisfies  $C_{inc}$  and not  $C_{exc}$  then
12:    Add  $a$  to  $\mathcal{L}$ 
13:  end if
14: end for
15: for each article  $a \in \mathcal{L}$  do
16:  Retrieve the full text of  $a$  and perform forward and backward citation chaining to
    identify additional articles
17:  for each new article  $b$  identified from citations do
18:    if  $b$  satisfies  $C_{inc}$  and not  $C_{exc}$  then
19:      Add  $b$  to  $\mathcal{L}$ 
20:    end if
21:  end for
22: end for
23: return  $\mathcal{L}$ 
  
```

3. Conceptual Foundations

This section establishes the theoretical foundation for understanding the roles of interpretability, explainability, and transparency in modern AI systems. Before examining specific methods and governance strategies, we clarify how these concepts differ, how they relate to fairness and accountability, and how they intersect with robustness and safety. Situating the discussion in both scholarly research and emerging regulatory frameworks helps build a shared vocabulary and highlights that explainability is not merely a technical exercise, but a human- and context-dependent requirement shaped by ethical norms, stakeholder needs, and societal expectations. Table 1 contains a summary of explanation types.

Table 1. Explanation types identified in regulatory guidance and the XAI literature. Each type corresponds to a distinct aspect of transparency and accountability.

Explanation Type	Purpose	Representative Sources
Rationale	Explain the reasoning behind a specific decision, including feature importance and statistical logic	ICO guidance [6]; Kim et al. [2]
Responsibility	Identify who is accountable for designing, deploying, and overseeing the model	ICO guidance [6]

Table 1. *Cont.*

Explanation Type	Purpose	Representative Sources
Data	Describe what data were used, how they were collected and processed, and how data quality and privacy were ensured	ICO guidance [6]; UNICEF policy [7]
Fairness	Outline measures taken to mitigate bias and ensure equitable outcomes	Deck et al. [8]; Brandl et al. [9]
Safety	Report performance metrics, robustness tests, security measures, and safety verification	Chen and Liu [4]; Madry et al. [10]
Impact	Discuss potential consequences for individuals and society, and engage stakeholders in evaluating impacts	ICO guidance [6]; Warren et al. [11]

3.1. Interpretability, Explainability and Transparency

Before discussing specific techniques, we clarify how key terms are used. Interpretability refers to the degree to which a person can understand an AI system's internal workings and predict its behavior in a given context. Explainability is a broader property that encompasses interpretability as well as the ability to convey reasons for model behavior and development processes in an accessible manner. Transparency refers to the openness of model design, training data, evaluation, and governance, allowing external actors to scrutinize and audit the system. These definitions draw on widely cited discussions of interpretability and model transparency [12,13]. Regulatory guidance, such as that from the UK Information Commissioner's Office, distinguishes six explanation types: rationale, responsibility, data, fairness, safety, and impact (each addressing different aspects of transparency and accountability) [6]. Moreover, interpretability exists on a spectrum: simple linear models are often considered intrinsically interpretable because their coefficients directly map to input features, whereas deep neural networks are typically opaque and rely on post-hoc interpretation methods. Scholars have also cautioned that claims about the fairness benefits of explanation may lack normative grounding and ignore power asymmetries [8,9]. A rigorous approach to interpretability must therefore grapple with normative commitments and stakeholder diversity.

3.2. Fairness, Accountability and Ethical Considerations

Fairness is a multifaceted concept that encompasses distributive, procedural, and contextual dimensions. Distributive fairness concerns equal outcomes across groups; procedural fairness concerns the fairness of the decision-making process; and contextual fairness recognizes the influence of social inequities. Formal fairness metrics, such as demographic parity, equalized odds, or equal opportunity, provide quantitative lenses but may conflict with one another. Recent work highlights that many XAI methods designed to report on fairness focus narrowly on procedural fairness and can be manipulated to present unfair models as fair [8]. A nuanced approach, therefore, requires broader ethical frameworks. The European Union's General Data Protection Regulation (GDPR) emphasizes lawfulness, fairness, and transparency in data processing, while the proposed AI Act introduces risk-based obligations. The ICO's guidance urges organizations to proactively disclose their use of AI, provide meaningful explanations, and assign responsibility for model oversight [6]. UNICEF's policy guidance on AI for children emphasizes the importance of age-appropriate explanations and protections for young users [7]. Throughout this paper, we refer to these frameworks when discussing design and deployment strategies.

In addition to formal metrics, documentation practices such as model cards and data statements help surface potential biases and support accountable AI development [14,15].

Recent research proposes fairness-aware training algorithms that integrate fairness penalties into objective functions to mitigate discrimination while maintaining performance [16,17]. Evaluation frameworks developed after 2020 offer systematic pipelines for measuring fairness across demographic groups and tasks, including graph neural networks and recommender systems [18,19]. These tools complement ethical guidelines by providing concrete methods for auditing and improving AI models.

3.3. *The Need for Robustness and Safety in Modern AI Systems*

Robustness refers to an AI system's ability to maintain performance under distributional shifts, noise, or adversarial attacks. The deep learning revolution has exposed vulnerabilities: neural networks can be fooled by imperceptible perturbations, poisoned training data, or hidden backdoors [20]. Surveys show that many organizations lack preparedness to secure their AI systems, and a significant fraction of cyberattacks involve data poisoning, model theft, or adversarial examples [4]. Robustness, therefore, spans natural robustness to shifts in data distribution, adversarial robustness against worst-case perturbations, and reliability under resource constraints or hardware faults. Robustness is intimately linked with interpretability and fairness: models that rely on spurious correlations may be both brittle and discriminatory, while adversarial training can increase robustness but alter feature importance, complicating explanations. Safety encompasses robustness, reliability, security, and the capacity to avoid harmful or unintended behaviors. We discuss these interactions in detail in Section 7.

3.4. *Conflicts and Synergies Among Interpretability, Fairness, and Robustness*

The relationships among interpretability, fairness, and robustness are multifaceted, exhibiting both tensions and complementarities that must be managed in practice. On one hand, interpretable models and explanation techniques can surface discriminatory correlations and spurious features, enabling developers to audit and mitigate unfair outcomes. Explanations may reveal that a classifier relies on sensitive attributes or highly correlated proxies, violating ethical expectations of distributive and procedural fairness [9]. Conversely, interventions designed to enforce fairness—such as constraining model outputs or reweighting training data—can alter a model's decision boundary and reduce its interpretability, because the resulting decision logic may be less aligned with human-intuitive features. Similarly, adversarial training aimed at improving robustness can encourage models to depend on subtle, human-imperceptible patterns, thereby decreasing the fidelity and usability of explanations. Empirical studies on explanation consistency show that co-training models for robustness and interpretability can partially offset these tensions: methods like Explanation Consistency Training and ensemble explanation alignment regularize models to produce stable attributions under perturbations while preserving accuracy and robustness [21,22]. Synergies also arise when robustness encourages reliance on human-perceptible features; for example, adversarially trained vision models often align saliency maps with object boundaries, enhancing both robustness and interpretability. These observations illustrate that fairness, robustness, and interpretability should not be pursued in isolation; instead, they require joint optimization and contextual ethical analysis to navigate trade-offs and leverage synergies.

4. Technical Approaches to Interpretability

Having established the conceptual foundations of interpretability, explainability, and transparency, we now turn to the methods for operationalizing these concepts in practice. A diverse toolbox of techniques has emerged, ranging from intrinsically interpretable models to post-hoc explanation strategies that attempt to open the black box. This section outlines

the primary classes of interpretability methods, highlighting their respective strengths, limitations, and appropriate use cases [23].

Before detailing individual techniques, Table 2 summarizes the principal classes of interpretability methods, highlighting their strengths, limitations, and representative references.

Table 2. Comparison of intrinsic and post-hoc interpretability methods. Intrinsic approaches are transparent by design, whereas post-hoc approaches generate explanations for models that are otherwise opaque.

Approach	Examples	Key Advantages	Limitations	References
Intrinsic	Linear models, decision trees, rule lists, generalized additive models	Transparent mapping from inputs to outputs; easy to audit	May sacrifice accuracy on complex tasks or unstructured data	Rudin [24]; Doshi-Velez and Kim [25]; Arrieta et al. [13]
Post-hoc (local)	LIME, SHAP, counterfactual explanations	Provide instance-specific explanations; model-agnostic	Explanations can be unstable and may not capture global logic	Ribeiro et al. [26]; Lundberg and Lee [27]
Post-hoc (global)	Feature attribution, saliency maps, layer-wise relevance	Offer global insights into model behavior	Often limited to specific architectures; may obscure causality	Gilpin et al. [28]; Roscher et al. [29]
Mechanistic	Circuit analysis, feature visualization, network dissection, SemanticLens	Reveal internal structures and functions, and scale to large multimodal models via semantic projection	Labour intensive and still under development for billion-parameter architectures	Muñoz et al. [30]; Dreyer et al. [31]

4.1. Intrinsic Interpretability

One class of methods seeks to design models that are interpretable by construction. Linear models, decision trees, rule lists, and scoring systems fall into this category. Their appeal lies in their simplicity: humans can trace the sequence of operations from inputs to outputs. However, these models may sacrifice accuracy when applied to high-dimensional or unstructured data. Recent advances explore richer intrinsically interpretable architectures, including generalized additive models with shape constraints, monotonic gradient boosting, and prototype-based networks. Rudin advocates for using such models in high-stakes domains, arguing that post-hoc explanations for black box models are inherently unfaithful. The trade-off between interpretability and performance is therefore domain-dependent; in areas with structured tabular data, interpretable models may perform comparably to black-box models, whereas tasks such as image classification often require deep networks and thus post-hoc methods.

4.2. Post-Hoc Explanation Methods

When black-box models are unavoidable, post-hoc techniques can approximate the reasoning behind their predictions. Local methods, such as Local Interpretable Model-agnostic Explanations (LIME) and SHapley Additive exPlanations (SHAP), approximate complex decision boundaries by fitting simple surrogate models around individual predictions. Gradient-based saliency maps, layer-wise relevance propagation, and integrated gradients attribute importance to input features by analyzing the network's derivatives. Counterfactual explanations generate hypothetical scenarios that change the model's output, thereby illuminating decision boundaries and suggesting actionable recourse. Other approaches visualize intermediate representations or synthesize human-readable concepts by clustering hidden activations. While powerful, these methods come with caveats: explanations

may be unstable under slight perturbations, susceptible to adversarial manipulation, or divergent across techniques. Mechanistic interpretability aims to delve deeper by analyzing network weights, circuits, and feature representations; progress in mechanistic approaches has been promising in smaller networks, but scaling them to frontier models remains a significant challenge. Ultimately, post-hoc tools should be used in conjunction with robust testing and human oversight.

To make these qualitative descriptions more concrete, we recall the mathematical foundations of several widely used post-hoc methods. In LIME, a complex classifier f is locally approximated by a simple linear surrogate $\hat{f}$ around a specific instance x . The surrogate is fit by minimizing a locally weighted loss over perturbed samples z drawn around x :

$$\hat{f}(z) = \beta_0 + \sum_{j=1}^p \beta_j z_j, \quad (1)$$

where z_j are binary indicators of interpretable features, and the coefficients β_j capture each feature's contribution. Weights decay with the distance between z and the original instance x .

SHAP explanations borrow from cooperative game theory to assign each feature i a Shapley value that represents its marginal contribution across all possible subsets of features:

$$\phi_i(f, x) = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|! (|F| - |S| - 1)!}{|F|!} [f(x_{S \cup \{i\}}) - f(x_S)], \quad (2)$$

where F is the full feature set and x_S denotes the input with only the subset S present. Integrated gradients provide another attribution method by accumulating gradients along a straight path from a baseline input x' to the input of interest x :

$$\text{IG}_i(x) = (x_i - x'_i) \int_0^1 \frac{\partial f(x' + \alpha(x - x'))}{\partial x_i} d\alpha. \quad (3)$$

These equations formalize how local surrogate models and attribution methods assign importance to input features, anchoring post-hoc explanations in quantitative measures. Recent work has extended these methods to structured data and multimodal inputs [30,32,33], motivating ongoing research into their stability and faithfulness.

Recent work demonstrates that mechanistic analysis can also elucidate how large language models implement specific functions. For example, Tak et al. show that autoregressive language models represent emotions in functionally localized subspaces and that targeted interventions on appraisal concepts can steer generated outputs [34]. Their study combines probing, causal mediation analysis, and intervention experiments to link emergent representations with psychological theory, illustrating that mechanistic insights can connect neural computations to human-interpretable concepts. Such findings suggest that mechanistic interpretability is maturing beyond toy models, enabling causal control and alignment in large models.

Beyond language models, mechanistic analyses are now being applied to large multimodal architectures that integrate text, vision, and audio. Techniques such as SemanticLens decompose activation patterns into semantically interpretable components across modalities, allowing researchers to trace how specific neurons or attention heads encode concepts like color, object categories, or sentiment in both textual and visual streams [31]. Other studies combine causal mediation analysis, network dissection, and probing across modalities to identify modality-specific circuits within vision-language transformers. These

investigations show that mechanistic tools can reveal shared or distinct pathways across different input modalities and can be used to diagnose misalignments or biases in multimodal models. Although scaling such methods to frontier models with billions of parameters remains computationally intensive, early results suggest that mechanistic interpretability is becoming a practical avenue for auditing and controlling complex systems beyond single-modality networks.

4.3. Human-Centered Evaluation and Trust

Interpretability methods must be evaluated with real users to ensure that explanations are meaningful and lead to appropriate trust. Systematic reviews demonstrate that explainability is inherently human-centered and that evaluating explanation quality requires experimental studies with target stakeholders. In clinical settings, for example, clear and concise explanations can increase clinicians' trust in AI recommendations, whereas overly complex or contradictory explanations may undermine trust. Public attitudes research reveals that while people value interpretability, many prioritize accuracy in high-stakes applications. Trust is therefore not unconditional: excessive trust in inaccurate models can be as harmful as skepticism toward reliable ones. Evaluation frameworks should measure not only subjective satisfaction but also whether explanations improve decision quality, calibrate trust, and align with ethical principles.

Recent human–subject experiments provide concrete evidence of these dynamics. A controlled study using explainable echo state networks showed that visual explanations significantly increase participants' trust and understanding of the model's decisions. Importantly, the effect was not moderated by age, gender, or prior experience, indicating that explanation-driven trust calibration can generalize across diverse users [35]. Such findings highlight the need to combine interpretability with transparent communication to foster appropriate reliance on AI systems.

5. Empirical Case Study: Balancing Interpretability, Performance, and Fairness

To complement our conceptual and policy survey, we undertook an empirical case study that demonstrates how interpretability, predictive performance, and fairness trade off against one another. Real-world applications, such as lending, hiring, and healthcare, have revealed how seemingly neutral models can inadvertently encode correlations between sensitive attributes and target outcomes, leading to disparate impacts on protected groups. Synthetic datasets offer a controlled environment for exploring these dynamics and illustrating patterns observed in practice. Our simulation loosely mirrors scenarios in which the protected attribute correlates with socioeconomic status, allowing us to test how different algorithms balance predictive power and fairness under controlled conditions [14,36]. All methods and code used to generate the dataset and evaluate the models are provided in a supplemental notebook to facilitate replication.

We generated a synthetic binary classification dataset with 2000 examples, 10 numerical features drawn from a multivariate normal distribution, and a binary sensitive attribute $A \in \{0,1\}$ that is moderately correlated with the ground-truth label Y . For each of five random seeds, we split the data into 80% training and 20% test sets. We trained three models: (1) a logistic regression classifier as a baseline intrinsically interpretable method, (2) a random forest with 100 trees representing a non-linear ensemble, and (3) a gradient boosting model (XGBoost) with default hyperparameters. Accuracy and F_1 score were computed to evaluate predictive performance on the test set.

Group-fairness was assessed using the demographic parity (DP) difference and the equalized odds (EO) difference (see Table 3). The DP difference measures the absolute difference in the proportion of positive predictions between the two sensitive groups:

$$\Delta_{DP} = \left| \Pr(\hat{Y} = 1 \mid A = 1) - \Pr(\hat{Y} = 1 \mid A = 0) \right|, \quad (4)$$

while the EO difference captures the maximum disparity in true positive and false positive rates across the groups:

$$\Delta_{EO} = \max_{y \in \{0,1\}} \left| \Pr(\hat{Y} = 1 \mid A = 1, Y = y) - \Pr(\hat{Y} = 1 \mid A = 0, Y = y) \right|. \quad (5)$$

Lower values of Δ_{DP} and Δ_{EO} indicate more equitable treatment of the sensitive groups. These metrics, commonly used in recent fairness research [37], are not simultaneously satisfiable in general and must be chosen based on contextual values [38].

Table 3. Mean performance and fairness metrics across five random seeds for logistic regression, random forest, and gradient boosting models on a synthetic dataset with a binary sensitive attribute. Higher accuracy and F_1 are better; lower DP and EO differences indicate fairer outcomes.

Model	Accuracy	F_1 Score	DP Difference	EO Difference
Logistic regression	0.787	0.784	0.057	0.034
Random forest	0.923	0.925	0.089	0.027
Gradient boosting	0.907	0.908	0.084	0.018

The case study highlights that no single model can simultaneously optimize interpretability, accuracy, and fairness. The simple logistic regression model is inherently interpretable but lags in predictive power; ensemble methods boost accuracy but can exacerbate group disparities and obscure decision logic. These findings align with contemporary fairness analyses, which show that group-fairness metrics may conflict and must be selected in accordance with stakeholder values and legal requirements [37,38]. They also emphasize the importance of standardized benchmarks and multi-objective optimization techniques that jointly consider accuracy, interpretability, and fairness when deploying AI systems in high-stakes domains.

It is essential to acknowledge that this case study is limited in scope and does not aim to provide universally generalizable conclusions. The synthetic dataset allows us to isolate correlations and illustrate trade-offs in a controlled environment. Still, real-world data often exhibit complex structural biases, high-dimensional interactions, and evolving distributions that this framework does not capture. Accordingly, the results should be interpreted as illustrative evidence of competing objectives rather than definitive guidance for deployment. Future work using real-world datasets is necessary to confirm whether the observed patterns hold in practice and to uncover additional nuances in fairness and robustness.

6. Control and Governance of AI Systems

Interpretability methods provide valuable insights into model behavior, but responsible AI also requires mechanisms that maintain human control. Governance structures, oversight processes, and stakeholder engagement ensure that AI systems are used appropriately and that decision makers remain accountable. The following section examines how human-in-the-loop designs, organizational policies, and context-aware explanations can align AI outputs with societal values.

6.1. Human-in-the-Loop and Oversight Mechanisms

Control refers to the ability of humans or institutions to guide, supervise, and correct AI behavior. Human-in-the-loop architectures retain a human decision maker who can accept, reject, or modify model outputs. These systems require clear interfaces that present model recommendations and explanations without overwhelming users. The ICO's transparency maxim emphasizes proactively disclosing AI use, providing truthful and timely explanations, and identifying responsible parties [6]. Accountability mechanisms assign responsibility for different stages of model development and deployment, ensure auditability, and enable redress for affected individuals. For example, process-based responsibility explanations should specify who collected the training data, designed the model, performed bias mitigation, and will conduct human reviews. Organizations should document model development through data fact sheets, model cards, and stakeholder impact assessments [39]. Oversight bodies, including regulators and ethics boards, play a crucial role in ensuring compliance with legal and ethical standards.

6.2. Contextual and Stakeholder-Aware Explanations

Different stakeholders require tailored explanations. The ICO and the Alan Turing Institute propose six explanation types: rationale, responsibility, data, fairness, safety, and impact [6]. Rationale explanations clarify why a model produced a specific outcome, including feature importance and statistical reasoning. Responsibility explanations identify who is accountable for the model's design and implementation. Data explanations describe the data used, how they were collected, processed, and protected. Fairness explanations outline steps taken to mitigate bias and ensure equitable outcomes. Safety explanations report performance metrics, robustness tests, and security measures. Impact explanations discuss potential consequences for individuals and society. Child-centered AI systems require age-appropriate language and graphic representations to communicate these aspects effectively [7]. Presenting layered explanations—beginning with high-level summaries and allowing users to delve deeper—can prevent information overload. Explanations should be offered as part of a dialogue, enabling questions and appeals. In regulated domains, explanations may need to meet specific legal requirements such as GDPR Article 22 or the EU AI Act's transparency obligations.

7. Robustness and Safety

Even the most interpretable and well-governed AI system can fail if it is brittle. Robustness and safety address a model's resilience to adversarial manipulation, distributional shifts, and operational hazards. This section surveys standard threat models, summarizes defense strategies, and discusses how to assess and ensure the reliability of AI systems in the face of uncertainty.

7.1. Adversarial Threats and Vulnerabilities

Deep learning models are vulnerable to a range of attacks. Evasion attacks add imperceptible perturbations to inputs to cause misclassification. Poisoning attacks manipulate training data to cause the learned model to behave maliciously. Backdoor attacks implant triggers that, when present, cause the model to output attacker-chosen labels. Surveys indicate that many organizations lack the knowledge to secure their AI systems, highlighting the need to address adversarial robustness throughout the AI lifecycle [4]. Threat models vary in the attacker's knowledge, including white box (complete understanding of the model), black box (limited to query access), and gray box (partial knowledge). Robustness also encompasses distributional shifts, such as changes in sensor calibration or population demographics, as well as hardware faults or resource constraints.

7.2. Defense Strategies

Defense strategies against adversarial attacks can be categorized into several types. Adversarial training augments the training set with adversarial examples, enabling the model to learn robust decision boundaries [10]. Certified defenses provide mathematical guarantees on model performance within specific perturbation norms using techniques such as interval-bound propagation or randomized smoothing. Input preprocessing removes adversarial noise through filtering or denoising, though adaptive attacks can circumvent such defenses. Ensemble and stochastic methods randomize model parameters or architectures to make attacks harder. Complementary to robustness is safety verification (e.g., formal methods that exhaustively search for errors or prove their absence in bounded regions of the input space). In addition, adversarial robustness interacts with interpretability: adversarial training can lead models to rely on human-perceptible features, potentially improving saliency maps. At the same time, some defenses may reduce interpretability by making gradients less informative [40]. Balancing robustness, accuracy, and interpretability remains an open research challenge. Table 4 summarizes defense strategies against adversarial attacks.

Table 4. Defense strategies against adversarial attacks. Each category outlines representative techniques, advantages, and limitations.

Category	Example Techniques	Advantages	Limitations
Adversarial training	Augmenting data with adversarial examples (Madry et al. [10])	Improves robustness to known attacks; can align saliency with human-perceptible features	Computationally expensive; may generalize poorly to unseen attacks
Certified defenses	Interval bound propagation; randomized smoothing (Carlini & Wagner [41])	Provide provable robustness guarantees within specified norms	Often limited to small networks or specific perturbation sizes; can reduce accuracy
Input preprocessing	JPEG compression; feature squeezing; denoising autoencoders (Papernot et al. [42])	Simple to implement; can mitigate some perturbations	Vulnerable to adaptive attacks; may harm clean accuracy
Ensemble and stochastic methods	Randomized smoothing; model ensembling (Ilyas et al. [40])	Increase unpredictability of model response to perturbations	Require multiple models or stochastic sampling; may still be circumvented
Formal verification	Reachability analysis; SMT solvers (Chen & Liu [4])	Provide exhaustive guarantees on behavior under bounded inputs	Scalability challenges for high-dimensional networks; often approximate

7.3. Safety, Reliability and Resilience

Safety extends beyond adversarial robustness to include reliability under uncertainty, protection of private and sensitive data, and resilience to unexpected events. Safety objectives should be defined along multiple dimensions: performance (accuracy and precision), reliability (faithful execution of intended functions), security (protection against unauthorized access), and robustness (resistance to perturbations). Safety assessments may use confusion matrices, receiver operating characteristic curves, calibration curves, and uncertainty estimates. Continuous monitoring for concept drift and model degradation is essential. Regulatory initiatives, such as the NIST AI Risk Management Framework and the NTIA Accountability Policy Report, emphasize the need for robust evaluations, incident reporting, and testbeds for piloting AI systems under controlled conditions [43,44]. Com-

plementing technical safeguards with organizational controls, such as role-based access, incident response plans, and independent audits, enhances resilience.

Beyond adversarial examples, distribution shifts and data incompleteness pose significant challenges to AI robustness. In real-world deployments, models trained on one domain may face shifted feature distributions or missing data in production. For example, medical imaging systems often encounter domain adaptation issues when scanning equipment or patient populations differ across hospitals, and autonomous vehicles must generalize from simulated environments to diverse road conditions. Recent work on generative models for fairness demonstrates that diffusion-based data augmentation can mitigate distribution shifts and improve performance and equity across histopathology, chest X-ray, and dermatology tasks [45]. In neuroimaging analysis, high-dimensional fMRI data are highly nonlinear and incomplete; a deep wavelet temporal-frequency attention factorization (Deep WTF) method reconstructs missing signals and assigns temporal-frequency weights, significantly enhancing robustness on autism spectrum disorder classification [46].

Traditional linear methods often struggle to capture their dynamic characteristics. Recently proposed temporal-frequency attention factorization methods [47] use temporal-frequency attention mechanisms to weight important features and reconstruct missing signals, significantly improving robustness and classification performance in Autism Spectrum Disorder classification tasks. Such approaches demonstrate that combining signal processing with deep learning can effectively address the challenges posed by high-dimensional, incomplete data. These methods suggest that robustness research must consider data heterogeneity, missingness, and domain shifts in addition to adversarial perturbations.

Robustness and interpretability are intertwined: adversarial training can guide models to rely on human-perceptible features, potentially improving explanation quality. In contrast, techniques that enforce consistent explanations under data perturbations can serve as robustness regularizers. Explanation Consistency Training (ECT) is one such approach; it encourages models to produce similar gradients and feature attributions when inputs are slightly altered, bridging semi-supervised learning with interpretability and improving both performance and explanation fidelity [21]. Similarly, ensemble architectures that align explanations across multiple sub-models can reduce variability in feature attributions and enhance trust [22]. Jointly optimizing adversarial robustness and explanation consistency remains an open research area, with early work suggesting that co-training for robustness and interpretability yields more resilient and transparent models.

The practical benefits of temporal-frequency attention factorization derive from its ability to capture multi-scale patterns in neural time series and to highlight clinically informative channels while suppressing noise. By assigning weights across both temporal segments and frequency bands, these methods focus the model's capacity on salient oscillatory rhythms, facilitating the reconstruction of missing signals and improving classification performance in autism spectrum disorder and related neurodevelopmental conditions. Compared with conventional linear models, this attention-based approach thus yields richer representations, enhances robustness under incomplete data, and can provide interpretable insights into which temporal and spectral regions drive diagnostic predictions [46,47].

8. Policy Landscape and Governance

Technical solutions alone cannot guarantee trustworthy AI; legal frameworks and ethical principles play a vital role in shaping how AI systems are developed and deployed. Governments, standards bodies, and civil society have proposed a range of policies and guidelines to strike a balance between innovation and fundamental rights. In this section, we outline key regulatory initiatives and normative frameworks that inform the design of

interpretable, controllable, and robust AI. Scholars of law and technology further argue that effective AI governance must connect technical safeguards with evolving data privacy statutes and civil-rights protections [48].

8.1. Regulatory Frameworks

Governments and standards bodies worldwide are crafting policies to ensure that AI systems align with societal values. The White House AI Action Plan calls for investing in AI interpretability, control, and robustness and recommends launching development programs and hackathons to advance these areas [1]. The European Union’s AI Act proposes a risk-based regulatory framework that requires transparency, human oversight, and documentation proportional to the AI system’s risk level. The GDPR establishes a right to meaningful information about automated decisions and mandates that personal data be processed lawfully, fairly, and transparently. The United Kingdom’s ICO provides practical guidance on transparency and explainability, emphasizing the disclosure of AI use, the provision of meaningful explanations, and the straightforward assignment of responsibility [6]. UNICEF’s policy guidance on AI and children emphasizes the importance of age-appropriate explanations, data minimization, and safeguarding children’s rights [7]. Standards organizations such as NIST have released guidance on explainable AI and the AI Risk Management Framework [43], while DARPA’s XAI program has spurred research on interpretable methods [49]. Together, these frameworks highlight the importance of documentation, auditing, and stakeholder engagement throughout the AI lifecycle. Table 5 adeptly consolidates all of these elements.

Table 5. Comparison of AI governance frameworks across selected regions. Each framework outlines principles for transparency, fairness, and accountability, with varying legal status and scope.

Region/Body	Framework	Key Principles	Legal Status
United States	AI Action Plan	Invests in interpretability, control, robustness; promotes innovation through testbeds and hackathons	Non-binding guidance
United States	AI Bill of Rights	Safe and effective systems, algorithmic discrimination protections, data privacy, notice and explanation, human alternatives	Executive policy blueprint; not codified
European Union	AI Act	Risk-based regulation; transparency, human oversight, documentation proportional to risk; prohibits certain practices	Pending legislation
United Kingdom	ICO Guidance	Transparency, meaningful explanations, accountability, data minimization	Regulatory guidance under GDPR
Global	GDPR	Lawfulness, fairness, transparency in data processing; right to explanation	Enforceable regulation
UNESCO	Recommendation on Ethics of AI	Human dignity, fairness, transparency, accountability, sustainability	Non-binding recommendation

Two recent U.S. initiatives illustrate different emphases within national AI policy. America's AI Action Plan, released under the Trump administration, frames AI as a strategic asset and prioritizes accelerating innovation, building AI infrastructure, and fostering international competitiveness. It calls for investments in interpretable, controllable, and robust AI systems through technology development programs, hackathons, and evaluation testbeds [1]. In contrast, the Biden administration's *Blueprint for an AI Bill of Rights* articulates five civil-rights-oriented principles: safe and effective systems, algorithmic discrimination protections, data privacy, notice and explanation, and human alternatives, that should guide the design and deployment of automated systems [50]. The Bill of Rights explicitly addresses algorithmic discrimination and data privacy, requiring that people be notified when computerized systems are used and have access to human fallback options. While both documents emphasize fairness, transparency, and accountability, the Action Plan leans toward deregulatory measures to spur innovation, whereas the Bill of Rights centers on human dignity, equity, and individual rights. Effective governance will likely require integrating the innovation-driven perspective of the Action Plan with the rights-based safeguards of the Bill of Rights and emerging international regulations [51,52].

8.2. Implementation Obstacles and Cross-Jurisdictional Coordination

While high-level principles for AI governance converge on transparency, fairness, and accountability, translating these principles into enforceable regulations faces several obstacles. First, there is no universally accepted set of technical standards for interpretability, auditing, or robustness; organizations struggle to measure compliance without standardized metrics or certification processes [53]. Second, policy enforcement is complicated by global data flows and inconsistent privacy laws: a system operating across borders must reconcile conflicting requirements on data sovereignty, consent, and algorithmic discrimination, and regulators must coordinate across jurisdictions to avoid regulatory arbitrage [48]. Third, sector-specific regulations and voluntary guidelines often lack mechanisms for monitoring and redress, leaving gaps between aspirational principles and actual practice. Addressing these challenges requires investment in interoperable technical standards, collaborative frameworks for cross-border governance, and institutional capacity to audit and enforce AI policies.

8.3. Ethical Principles and Human Rights

Beyond compliance, AI governance should be grounded in ethical principles. UNESCO's Recommendation on the Ethics of Artificial Intelligence emphasizes human dignity, fairness, transparency, accountability, and environmental sustainability [54,55]. Philosophers such as Floridi argue for placing human values and rights at the center of AI design and caution against technological determinism. The principle of non-discrimination requires that AI systems avoid disparate impacts on protected groups, while the principle of beneficence seeks to maximize societal benefit. The right to explanation, articulated by Wachter and colleagues, positions interpretability within a broader framework of procedural justice [56]. Operationalizing these principles requires multidisciplinary collaboration among computer scientists, ethicists, legal scholars, domain experts, and affected communities, who must co-design and evaluate AI systems throughout their lifecycles [57].

9. Challenges and Research Directions

The preceding sections have highlighted significant advances in interpretability, control, and robustness, yet many open problems persist. Here, we synthesize the critical challenges that the research community must address and propose directions for future work to ensure that AI systems remain transparent, fair, and safe as they evolve.

Before enumerating specific challenges, it is important to clarify how interpretability should be measured. Beyond subjective impressions, researchers have developed quantitative metrics such as *explanation fidelity*—the degree to which an explanation accurately reflects the model’s actual decision logic—*explanation sufficiency*, which assesses whether the features highlighted in an explanation are sufficient to reproduce the prediction, and *user comprehension*, which evaluates whether the intended audience understands and can act on the explanation [33,58]. Recent frameworks, such as XAI-Eval, operationalize these metrics across diverse data modalities and provide standardized benchmarks. Combining these metrics with human-subject studies yields a more holistic assessment of explanation quality and informs the design of interpretability methods that are both faithful and usable [2].

Despite rapid progress, significant challenges remain. First, conceptual ambiguity around interpretability hampers clear expectations and evaluation; future work should develop standardized taxonomies and metrics that capture both statistical fidelity and human comprehension [33,58]. Second, scalable mechanistic interpretability techniques are needed to analyze large, multimodal models without oversimplification. Third, integrating interpretability with robustness presents trade-offs: adversarial training may improve robustness but alter feature importance, while some explanation methods can leak information that attackers exploit. Fourth, fairness remains contested; XAI must grapple with different notions of fairness and avoid becoming a veneer for unjust systems [9]. Fifth, explanations must be tailored to diverse audiences and socio-cultural contexts, including children and marginalized communities [2,11]. Finally, policy frameworks must strike a balance between innovation and safeguards, ensuring that regulations keep pace with technological advances without stifling beneficial research. Emerging surveys on large language models for XAI highlight opportunities to leverage generative models as explanatory tools while cautioning that these systems introduce new risks [59,60].

Our empirical case study relied on a synthetic dataset to illustrate trade-offs among interpretability, predictive performance, and fairness. Although synthetic data allow researchers to control correlations and isolate conceptual effects, they do not capture the complex structural biases, high-dimensional interactions, and evolving distributions present in real-world domains. Consequently, the observed patterns may not generalize to applications in sectors such as healthcare, finance, or education, where sensitive attributes interact with socioeconomic factors, medical histories, or market dynamics. Future research should validate these findings on curated real-world datasets from high-impact domains and develop benchmarks that reflect domain-specific constraints and structural biases [14,36]. Such work will enable more robust evaluation of interpretability, fairness, and robustness under authentic conditions and inform the design of mitigation strategies in practice.

10. Conclusions

Artificial intelligence is poised to reshape society, but realizing its benefits responsibly requires sustained investments in interpretability, control, and robustness. The opacity of frontier AI systems undermines trust, adversarial vulnerabilities expose models to manipulation, and unfair outcomes can exacerbate societal inequities. In this review, we synthesized technical advances and policy initiatives, emphasizing that interpretability is multifaceted and human-centered, that control depends on human oversight and governance structures, and that robustness and safety demand both technical and organizational safeguards. We also reflected on emerging topics such as mechanistic interpretability, human-subject experiments on trust calibration, joint optimization of robustness and interpretability, challenges posed by distribution shifts and incomplete data, and comparative analyses of AI governance frameworks. An empirical case study illustrated how interpretability, predictive performance, and fairness intertwine in practice, showing that

interpretable models can deliver equitable outcomes but may sacrifice predictive accuracy. In contrast, complex ensembles improve performance at the expense of fairness and transparency.

Looking forward, several technical and organizational strategies can guide the development of trustworthy AI. Multi-objective optimization offers a principled way to balance accuracy, interpretability, and fairness by incorporating fairness penalties or constraints into training objectives and exploring Pareto frontiers. Co-designing robustness and interpretability, for example, through explanation consistency training or adversarial training on human-perceptible features, can simultaneously enhance resilience and explanation quality. Generative augmentation techniques and methods, such as deep wavelet temporal frequency attention factorization, address distribution shifts and high-dimensional incomplete data, improving robustness and fairness under real-world conditions. Evaluation standards for interpretability are maturing, with metrics such as explanation fidelity, sufficiency, and user comprehension now complementing qualitative user studies.

Researchers should develop quantitative metrics for explanation fidelity and usefulness, complement qualitative user studies, and pursue scalable mechanistic analyses. Practitioners are encouraged to employ a suite of evaluation tools—including calibration curves, confusion matrices, receiver operating characteristic curves, and group fairness metrics—to assess performance across multiple dimensions and to provide layered explanations tailored to diverse stakeholders. Policymakers should harmonize innovation-oriented policies with rights-based frameworks, ensuring that regulatory guidance reflects both the need for AI competitiveness and the imperative to protect civil rights, privacy, and human dignity. A comparative policy analysis underscores that AI governance frameworks vary widely across regions and sectors; aligning regulations with rapid technical advances will require cross-jurisdictional dialogue and continuous updates to enforce principles of fairness, accountability, and transparency in practice.

Funding: This research received no external funding.

Data Availability Statement: The original contributions presented in this study are included in the article. Further inquiries can be directed to the author.

Conflicts of Interest: The author declares no conflicts of interest.

References

1. Hine, E.; Floridi, L. United States · Winning the AI Race? The US AI Action Plan in Context. *J. Law Regul.* **2025**, *2*, 274–278. [\[CrossRef\]](#)
2. Kim, J.; Maathuis, H.; Sent, D. Human-centered evaluation of explainable AI applications: A systematic review. *Front. Artif. Intell.* **2024**, *7*, 1456486. [\[CrossRef\]](#)
3. Baker, S.; Xiang, W. Explainable AI is Responsible AI: How Explainability Creates Trustworthy and Socially Responsible Artificial Intelligence. *arXiv* **2023**, arXiv:2312.01555. [\[CrossRef\]](#)
4. Chen, P.Y.; Liu, S. Holistic Adversarial Robustness of Deep Learning Models. *Proc. AAAI Conf. Artif. Intell.* **2023**, *37*, 15411–15420. [\[CrossRef\]](#)
5. Leon, M. GPT-5 and open-weight large language models: Advances in reasoning, transparency, and control. *Inf. Syst.* **2026**, *136*, 102620. [\[CrossRef\]](#)
6. Leslie, D. Explaining Decisions Made with AI. *SSRN Electron. J.* **2020**. [\[CrossRef\]](#)
7. Liu, S.; Ding, W. Artificial intelligence for children: UNICEF's policy guidance and beyond. *Child. Soc.* **2024**, *39*, 374–382. [\[CrossRef\]](#)
8. Deck, L.; Schoeffer, J.; De-Arteaga, M.; Kühl, N. A Critical Survey on Fairness Benefits of Explainable AI. In *FAcT '24: Proceedings of the 2024 ACM Conference on Fairness Accountability and Transparency*; ACM: New York, NY, USA, 2024; pp. 1579–1595. [\[CrossRef\]](#)

9. Brandl, S.; Bugliarello, E.; Chalkidis, I. On the Interplay between Fairness and Explainability. In *Proceedings of the 4th Workshop on Trustworthy Natural Language Processing (TrustNLP 2024)*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2024; pp. 94–108. [\[CrossRef\]](#)
10. Reddy Kandula, S. Adversarial Resilience in Deep Learning: Challenges, Defense Mechanisms, and Future Directions. *J. Recent Trends Comput. Sci. Eng.* **2025**, *13*, 1–14. [\[CrossRef\]](#)
11. Warren, G.; Shklovski, I.; Augenstein, I. Show Me the Work: Fact-Checkers' Requirements for Explainable Automated Fact-Checking. In *CHI '25: Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*; ACM: New York, NY, USA, 2025; pp. 1–21. [\[CrossRef\]](#)
12. Shivadekar, S. *Cognitive Artificial Intelligence for Health and Climate: Deep Models, Interpretability, and Decision Support*; Deep Science Publishing: New Delhi, India, 2025. [\[CrossRef\]](#)
13. Kabir, S.; Hossain, M.S.; Andersson, K. A Review of Explainable Artificial Intelligence from the Perspectives of Challenges and Opportunities. *Algorithms* **2025**, *18*, 556. [\[CrossRef\]](#)
14. Pushkarna, M.; Zaldivar, A.; Kjartansson, O. Data Cards: Purposeful and Transparent Dataset Documentation for Responsible AI. In *FACCT '22: Proceedings of the 2022 ACM Conference on Fairness Accountability and Transparency*; ACM: New York, NY, USA, 2022; pp. 1776–1826. [\[CrossRef\]](#)
15. Hutchinson, B.; Smart, A.; Hanna, A.; Denton, R.; Greer, C.; Kjartansson, O.; Barnes, P.; Mitchell, M. Towards Accountability for Machine Learning Datasets: Practices from Software Engineering and Infrastructure. In *FACCT '21: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*; ACM: New York, NY, USA, 2021; pp. 560–575. [\[CrossRef\]](#)
16. Nathim, K.W.; Hameed, N.A.; Salih, S.A.; Taher, N.A.; Salman, H.M.; Chornomordenko, D. Ethical AI with Balancing Bias Mitigation and Fairness in Machine Learning Models. In *Proceedings of the 2024 36th Conference of Open Innovations Association (FRUCT)*; IEEE: Piscataway, NJ, USA, 2024; pp. 797–807. [\[CrossRef\]](#)
17. Bin, C.; Liu, W.; Zhang, F.; Chang, L.; Gu, T. FairCoRe: Fairness-aware Recommendation through Counterfactual Representation Learning. *IEEE Trans. Knowl. Data Eng.* **2025**, *37*, 4049–4062. [\[CrossRef\]](#)
18. Lalor, J.P.; Abbasi, A.; Oketch, K.; Yang, Y.; Forsgren, N. Should Fairness be a Metric or a Model? A Model-based Framework for Assessing Bias in Machine Learning Pipelines. *ACM Trans. Inf. Syst.* **2024**, *42*, 1–41. [\[CrossRef\]](#)
19. Chen, A.; Rossi, R.A.; Park, N.; Trivedi, P.; Wang, Y.; Yu, T.; Kim, S.; Dernoncourt, F.; Ahmed, N.K. Fairness-Aware Graph Neural Networks: A Survey. *ACM Trans. Knowl. Discov. Data* **2024**, *18*, 1–23. [\[CrossRef\]](#)
20. Kats, J. Machine Learning Detection of IPKVM Exploitation in Online Exam Environments. In *Proceedings of the 2025 IEEE Opportunity Research Scholars Symposium (ORSS)*; IEEE: Piscataway, NJ, USA, 2025; pp. 1–5. [\[CrossRef\]](#)
21. Han, T.; Tu, W.W.; Li, Y.F. Explanation Consistency Training: Facilitating Consistency-Based Semi-Supervised Learning with Interpretability. *Proc. AAAI Conf. Artif. Intell.* **2021**, *35*, 7639–7646. [\[CrossRef\]](#)
22. Watson, M.; Awwad Shiekh Hasan, B.; Al Moubayed, N. Using model explanations to guide deep learning models towards consistent explanations for EHR data. *Sci. Rep.* **2022**, *12*, 19899. [\[CrossRef\]](#) [\[PubMed\]](#)
23. DeSimone, H. Explainable AI: The Quest for Transparency in Business and Beyond. In *Proceedings of the 2024 7th International Conference on Information and Computer Technologies (ICICT)*; IEEE: Piscataway, NJ, USA, 2024; pp. 532–538. [\[CrossRef\]](#)
24. David, R.; Shankar, H.; Kura, P.; Kowtarapu, K.; S, U.M.; Karkuzhali, S. Advancement in Explainable AI: Bringing Transparency and Interpretability to Machine Learning Models for Use in High-Stakes Decisions. In *Proceedings of the 2025 International Conference on Emerging Smart Computing and Informatics (ESCI)*; IEEE: Piscataway, NJ, USA, 2025; pp. 1–6. [\[CrossRef\]](#)
25. Kruschel, S.; Hambauer, N.; Weinzierl, S.; Zilker, S.; Kraus, M.; Zschech, P. Challenging the Performance-Interpretability Trade-Off: An Evaluation of Interpretable Machine Learning Models. *Bus. Inf. Syst. Eng.* **2025**. [\[CrossRef\]](#)
26. Nizette, F.; Hammedi, W.; van Riel, A.C.; Steils, N. Why should I trust you? Influence of explanation design on consumer behavior in AI-based services. *J. Serv. Manag.* **2024**, *36*, 50–74. [\[CrossRef\]](#)
27. Kakhani, N.; Taghizadeh-Mehrjardi, R.; Omarzadeh, D.; Ryo, M.; Heiden, U.; Scholten, T. Towards Explainable AI: Interpreting Soil Organic Carbon Prediction Models Using a Learning-Based Explanation Method. *Eur. J. Soil Sci.* **2025**, *76*, e70071. [\[CrossRef\]](#)
28. ŞAHİN, E.; Arslan, N.N.; Özdemir, D. Unlocking the black box: An in-depth review on interpretability, explainability, and reliability in deep learning. *Neural Comput. Appl.* **2024**, *37*, 859–965. [\[CrossRef\]](#)
29. Garg, P.; Sharma, M.K.; Kumar, P. Transparency in Diagnosis: Unveiling the Power of Deep Learning and Explainable AI for Medical Image Interpretation. *Arab. J. Sci. Eng.* **2025**, *50*, 15751–15767. [\[CrossRef\]](#)
30. Guleria, P.; Sood, M. Explainable AI and machine learning: Performance evaluation and explainability of classifiers on educational data mining inspired career counseling. *Educ. Inf. Technol.* **2022**, *28*, 1081–1116. [\[CrossRef\]](#)
31. Dreyer, M.; Berend, J.; Labarta, T.; Vielhaben, J.; Wiegand, T.; Lapuschkin, S.; Samek, W. Mechanistic understanding and validation of large AI models with SemanticLens. *Nat. Mach. Intell.* **2025**, *7*, 1572–1585. [\[CrossRef\]](#)
32. Agarwal, C.; Queen, O.; Lakkaraju, H.; Zitnik, M. Evaluating explainability for graph neural networks. *Sci. Data* **2023**, *10*, 144. [\[CrossRef\]](#) [\[PubMed\]](#)

33. Nauta, M.; Trienes, J.; Pathak, S.; Nguyen, E.; Peters, M.; Schmitt, Y.; Schlötterer, J.; van Keulen, M.; Seifert, C. From Anecdotal Evidence to Quantitative Evaluation Methods: A Systematic Review on Evaluating Explainable AI. *ACM Comput. Surv.* **2023**, *55*, 1–42. [\[CrossRef\]](#)
34. Tak, A.N.; Banayeeanzade, A.; Bolourani, A.; Kian, M.; Jia, R.; Gratch, J. Mechanistic Interpretability of Emotion Inference in Large Language Models. In *Proceedings of the Findings of the Association for Computational Linguistics: ACL 2025*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2025; pp. 13090–13120. [\[CrossRef\]](#)
35. Hao, S.; Teng, F.; Hou, R.; Zhang, L.; Wu, H.; Qi, J. Explainable AI and echo state networks calibrate trust in human-machine interaction. *Sci. Rep.* **2026**, *16*, 1189. [\[CrossRef\]](#)
36. Marusich, L.R.; Files, B.T.; Bancilhon, M.; Rawal, J.C.; Raglin, A., Trust Calibration for Joint Human/AI Decision-Making in Dynamic and Uncertain Contexts. In *Artificial Intelligence in HCI*; Springer Nature: Cham, Switzerland, 2025; pp. 106–120. [\[CrossRef\]](#)
37. Rabonato, R.T.; Berton, L. A systematic review of fairness in machine learning. *AI Ethics* **2024**, *5*, 1943–1954. [\[CrossRef\]](#)
38. Ashurst, C.; Weller, A. Fairness Without Demographic Data: A Survey of Approaches. In *EAAMO '23: Proceedings of the Equity and Access in Algorithms, Mechanisms, and Optimization*; ACM: New York, NY, USA, 2023; pp. 1–12. [\[CrossRef\]](#)
39. Leon, M. Cognitive mapping variants and their training algorithms. *Comput. Sci. Rev.* **2026**, *59*, 100862. [\[CrossRef\]](#)
40. Remtulla, R.; Samet, A.; Kulbay, M.; Akdag, A.; Hocini, A.; Volniansky, A.; Kahn Ali, S.; Qian, C.X. A Future Picture: A Review of Current Generative Adversarial Neural Networks in Vitreoretinal Pathologies and Their Future Potentials. *Biomedicine* **2025**, *13*, 284. [\[CrossRef\]](#)
41. Zühlke, M.M.; Kudenko, D. Adversarial Robustness of Neural Networks from the Perspective of Lipschitz Calculus: A Survey. *ACM Comput. Surv.* **2025**, *57*, 1–41. [\[CrossRef\]](#)
42. Chen, J.; Yan, H.; Liu, B.; Zhao, S.; Chen, X.; Li, Z.; Xu, H. Erasing backdoor of deep neural networks using neural perturbation-based attention distillation. *Int. J. Model. Simul. Sci. Comput.* **2025**, *16*. [\[CrossRef\]](#)
43. Tabassi, E. *Artificial Intelligence Risk Management Framework (AI RMF 1.0)*; National Institute of Standards and Technology: Gaithersburg, MD, USA, 2023. [\[CrossRef\]](#)
44. Novelli, C.; Taddeo, M.; Floridi, L. Accountability in artificial intelligence: What it is and how it works. *AI Soc.* **2023**, *39*, 1871–1882. [\[CrossRef\]](#)
45. Ktena, I.; Wiles, O.; Albuquerque, I.; Rebuffi, S.A.; Tanno, R.; Roy, A.G.; Azizi, S.; Belgrave, D.; Kohli, P.; Cemgil, T.; et al. Generative models improve fairness of medical classifiers under distribution shifts. *Nat. Med.* **2024**, *30*, 1166–1173. [\[CrossRef\]](#)
46. Wang, F.; Ke, H.; Cai, C. Deep Wavelet Self-Attention Non-negative Tensor Factorization for non-linear analysis and classification of fMRI data. *Appl. Soft Comput.* **2025**, *182*, 113522. [\[CrossRef\]](#)
47. Ke, H.; Wang, F.; Bi, H.; Ma, H.; Wang, G.; Yin, B. Unsupervised deep frequency-channel attention factorization to non-linear feature extraction: A case study of identification and functional connectivity interpretation of Parkinson's disease. *Expert Syst. Appl.* **2024**, *243*, 122853. [\[CrossRef\]](#)
48. Mbah, G.O. Data privacy in the era of AI: Navigating regulatory landscapes for global businesses. *Int. J. Sci. Res. Arch.* **2024**, *13*, 2040–2058. [\[CrossRef\]](#)
49. Chamola, V.; Hassija, V.; Sulthana, A.R.; Ghosh, D.; Dhingra, D.; Sikdar, B. A Review of Trustworthy and Explainable Artificial Intelligence (XAI). *IEEE Access* **2023**, *11*, 78994–79015. [\[CrossRef\]](#)
50. Weaver, J.F. AI Bill of Rights and creative lawmaking. In *Research Handbook on the Law of Artificial Intelligence*; Edward Elgar Publishing: Cheltenham, UK, 2025; pp. 36–56. [\[CrossRef\]](#)
51. Blumenthal-Barby, J. An AI Bill of Rights: Implications for Health Care AI and Machine Learning—A Bioethics Lens. *Am. J. Bioeth.* **2022**, *23*, 4–6. [\[CrossRef\]](#)
52. Hine, E.; Floridi, L. Artificial intelligence with American values and Chinese characteristics: A comparative analysis of American and Chinese governmental AI policies. *AI Soc.* **2022**, *39*, 257–278. [\[CrossRef\]](#)
53. Batool, A.; Zowghi, D.; Bano, M. AI governance: A systematic literature review. *AI Ethics* **2025**, *5*, 3265–3279. [\[CrossRef\]](#)
54. Leon, M. The Escalating AI's Energy Demands and the Imperative Need for Sustainable Solutions. *Wseas Trans. Syst.* **2024**, *23*, 444–457. [\[CrossRef\]](#)
55. Blanchard, A.; Taddeo, M. The Ethics of Artificial Intelligence for Intelligence Analysis: A Review of the Key Challenges with Recommendations. *Digit. Soc.* **2023**, *2*, 12. [\[CrossRef\]](#)
56. Bayamlıoğlu, E. The right to contest automated decisions under the General Data Protection Regulation: Beyond the so-called “right to explanation”. *Regul. Gov.* **2021**, *16*, 1058–1078. [\[CrossRef\]](#)
57. Napoles, G. Prolog-based agnostic explanation module for structured pattern classification. *Inf. Sci.* **2023**, *622*, 1196–1227. [\[CrossRef\]](#)
58. Agrawal, K.; El Shawi, R.; Ahmed, N. XAI-Eval: A framework for comparative evaluation of explanation methods in healthcare. *Digit. Health* **2025**, *11*, 1–16. [\[CrossRef\]](#)

59. Leon, M. Generative Artificial Intelligence and Prompt Engineering: A Comprehensive Guide to Models, Methods, and Best Practices. *Adv. Sci. Technol. Eng. Syst. J.* **2025**, *10*, 1–11. [[CrossRef](#)]
60. Bilal, A.; Ebert, D.; Lin, B. LLMs for Explainable AI: A Comprehensive Survey. *arXiv* **2025**, arXiv:2504.00125. [[CrossRef](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.