Article

Achieving Explainable, Scalable, and Robust Machine Learning for Real-World Applications

Yong Nuan and Zhihao Ru [*]

Posted Date: 30 June 2025

doi: 10.20944/preprints202506.2515.v1

Keywords: Explainable AI; Interpretability; Scalability; Robustness; Adversarial Learning; Model Compression; Feature Attribution; Regularization; Modular Architectures; Distributional Robustness; Trustworthy Machine Learning; Multi-Objective Optimization

Article

Achieving Explainable, Scalable, and Robust Machine Learning for Real-World Applications

Yong Nuan and Zhihao Ru *

Department of Computer Science, Chinese University of Hong Kong, Hong Kong; yong.nuan@cuhk.edu.hk
* Correspondence: hihao.ru@cuhk.edu.hk

Abstract: The increasing deployment of machine learning systems in high-stakes and resource-constrained environments has accentuated the necessity for models that are simultaneously explainable, scalable, and robust. While each of these desiderata has been extensively studied in isolation, their integration remains a critical open challenge due to inherent trade-offs and complex interactions. This paper presents a comprehensive framework that unifies theoretical foundations, methodological advances, and empirical evaluations to address these intertwined objectives. We formalize explainability through function decomposition and feature attribution, characterize scalability in terms of computational efficiency and statistical generalization, and define robustness via distributional and adversarial perturbations. Our survey of contemporary methods reveals a rich design space including regularization techniques, modular architectures, and robust optimization paradigms that can be systematically combined to achieve balanced performance. Extensive experiments across diverse datasets demonstrate Pareto-optimal trade-offs and highlight practical considerations for model selection and deployment. Finally, we discuss ethical implications, contextual constraints, and future research directions aimed at developing trustworthy machine learning systems that align with human values and operational demands.

Keywords: explainable AI; interpretability; scalability; robustness; adversarial learning; model compression; feature attribution; regularization; modular architectures; distributional robustness; trustworthy machine learning; multi-objective optimization

1. Introduction

The rapid expansion of machine learning (ML) applications across diverse domains—ranging from healthcare and finance to autonomous systems and scientific discovery—has underscored the urgent need for models that are not only accurate but also explainable, scalable, and robust [1]. While modern machine learning algorithms, particularly deep learning models, have demonstrated remarkable performance on a wide range of tasks, their widespread adoption is increasingly impeded by a lack of interpretability, limited scalability to large and heterogeneous data, and vulnerability to adversarial attacks, distributional shifts, and noise. These limitations pose significant challenges in high-stakes and safety-critical applications, where trust, accountability, and generalization are essential [2]. In traditional statistical modeling, the emphasis was often on simplicity, interpretability, and statistical guarantees [3]. However, the advent of complex neural architectures and massive datasets has shifted the focus towards empirical performance. As a consequence, many state-of-the-art models function as "black boxes", offering little insight into the decision-making processes that drive their predictions [4]. This opacity undermines trust, limits regulatory compliance, and hampers human-machine collaboration. To mitigate these concerns, the subfield of Explainable Artificial Intelligence (XAI) has emerged, seeking to develop techniques that illuminate the inner workings of complex models. These techniques include feature attribution methods, surrogate modeling, counterfactual explanations, and concept-based interpretability frameworks, each aiming to bridge the gap between high performance and human understanding [5]. Simultaneously, the scalability of

doi:10.20944/preprints202506.2515.v1

ML models remains a pressing issue as datasets continue to grow in size and complexity. Classical learning algorithms often suffer from prohibitive computational or memory requirements when applied to web-scale or real-time environments. Furthermore, model training and inference pipelines must contend with heterogeneous data modalities, non-stationary data distributions, and privacy constraints [6]. Advances in distributed learning, model compression, and efficient architecture design have offered partial solutions, but scalability remains a bottleneck in many practical deployments [7]. The challenge lies in designing methods that can learn from massive data streams, adapt in real time, and deliver performance guarantees with manageable resource footprints. Robustness, the third cornerstone of this paper, refers to the model's resilience to perturbations in input data or environmental conditions [3]. This includes robustness to adversarial examples, noisy or incomplete data, out-of-distribution samples, and domain shifts [8]. Robust models are particularly critical in adversarial or dynamic environments, such as cybersecurity, medical diagnostics, and autonomous navigation. Despite progress in robust optimization, certified defenses, and adversarial training, existing methods often trade off robustness with accuracy or interpretability, leading to fragile and brittle systems. A central question in contemporary machine learning research is how to balance these three pillars—explainability, scalability, and robustness—without sacrificing predictive performance [9]. In this paper, we aim to provide a unified perspective on the design and analysis of machine learning systems that are explainable, scalable, and robust. We begin by reviewing foundational concepts and formal definitions that underlie each of these desiderata [10]. We then survey recent progress in techniques that address these challenges, highlighting both synergistic approaches and inherent trade-offs. For example, while model distillation can aid interpretability and reduce model size, it may also introduce vulnerabilities. Similarly, sparsity-based approaches may enhance both robustness and transparency but at the cost of reduced expressiveness [11]. Our contributions are threefold. First, we propose a conceptual framework that systematically integrates explainability, scalability, and robustness, identifying shared principles and design patterns [12]. Second, we present a taxonomy of methods that have demonstrated effectiveness in one or more of these dimensions, analyzing their theoretical guarantees and empirical performance across benchmark tasks. Third, we introduce a set of evaluation metrics and standardized protocols that can be used to assess model performance beyond mere accuracy, fostering the development of holistic and trustworthy machine learning systems [13]. Ultimately, our goal is to move beyond the current paradigm of accuracy-centric model evaluation and advocate for a new generation of learning systems that are not only powerful but also transparent, efficient, and resilient. By illuminating the connections between explainability, scalability, and robustness[14,15], we hope to chart a path forward for building machine learning models that align more closely with human values and societal needs.

2. Theoretical Foundations

In order to systematically design machine learning models that are explainable, scalable, and robust, it is essential to ground our discussion in rigorous mathematical formalism. Let $\mathcal{X} \subseteq \mathbb{R}^d$ denote the input space, and $\mathcal{Y}$ denote the output space, where for classification tasks we typically have $\mathcal{Y} = \{1, 2, \ldots, C\}$, and for regression tasks, $\mathcal{Y} \subseteq \mathbb{R}$ [16]. A machine learning model is a function $f_\theta : \mathcal{X} \to \mathcal{Y}$ parameterized by $\theta \in \Theta$, trained to minimize a loss function $\mathcal{L} : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+$ over a data distribution $\mathcal{D}$ [17].

2.1. Explainability as a Function Decomposition Problem

Explainability can be viewed through the lens of function decomposition [18]. Given a learned model $f_\theta(x)$, we seek an interpretable approximation $\hat{f}(x)$ such that for a relevant set of inputs $x \in \mathcal{X}$, the approximation error $\|f_\theta(x) - \hat{f}(x)\|$ is small under some norm, typically L_2 or L_1 [19]. One common approach is to represent $\hat{f}(x)$ as a linear function in a local neighborhood:

$$\hat{f}(x) = w^\top x + b, \quad \text{for } x \in \mathcal{B}_\epsilon(x_0),$$

where $\mathcal{B}_\epsilon(x_0)$ denotes a ball of radius ϵ around some point of interest x_0. This formulation underpins local surrogate methods such as LIME and SHAP. The key theoretical question becomes: under what assumptions on f_θ and $\mathcal{D}$ does such a linear surrogate $\hat{f}$ yield faithful explanations?

2.2. Scalability and Statistical Generalization

Scalability is often conflated with computational efficiency, but it also encompasses statistical considerations [20]. Let n denote the number of training samples and d the input dimension. Traditional VC-dimension bounds imply that the generalization error depends on model complexity and sample size [21]. However, in the overparameterized regime where $d \gg n$, classical theory fails to explain empirical success [22]. Modern generalization bounds invoke notions such as Rademacher complexity, uniform stability, and norm-based capacity control. For instance, for a model class $\mathcal{F}$ and loss ℓ, we have the bound:

$$\mathbb{E}_{\mathcal{D}}\left[\sup_{f\in\mathcal{F}}\left|\mathbb{E}_{(x,y)\sim\mathcal{D}}[\ell(f(x),y)] - \frac{1}{n}\sum_{i=1}^{n}\ell(f(x_i),y_i)\right|\right] \leq \mathfrak{R}_n(\mathcal{F}) + O\left(\sqrt{\frac{1}{n}}\right),$$

where $\mathfrak{R}_n(\mathcal{F})$ is the empirical Rademacher complexity [23]. Scalability, then, entails constructing function classes $\mathcal{F}$ with low complexity relative to their expressive power, possibly via sparsity, low-rank structure, or modular design.

2.3. Robustness under Distributional Shifts

Robustness can be formalized as the stability of f_θ under perturbations to the input or the data distribution. Adversarial robustness focuses on perturbations within some ℓ_p-ball, i.e., $\|x' - x\|_p \leq \delta$. Distributional robustness, on the other hand, generalizes this notion by considering worst-case expected loss over an uncertainty set $\mathcal{P}$ of probability distributions:

$$\sup_{\mathbb{Q}\in\mathcal{P}} \mathbb{E}_{(x,y)\sim\mathbb{Q}}[\ell(f(x),y)].$$

Common choices for $\mathcal{P}$ include Wasserstein balls or f-divergence balls centered at the empirical distribution. Solving the corresponding minimax problem,

$$\min_{f\in\mathcal{F}} \sup_{\mathbb{Q}\in\mathcal{P}} \mathbb{E}_{\mathbb{Q}}[\ell(f(x),y)],$$

leads to models that are less sensitive to shifts and adversarial manipulation [24]. The duality between robustness and regularization further implies that robust optimization often introduces implicit inductive biases, which may interact non-trivially with explainability.

2.4. Illustration: Trade-off Surface

To visually encapsulate the interactions between the three desiderata—explainability, scalability, and robustness—we illustrate a conceptual trade-off surface [25]. While not exhaustive, this diagram provides intuition for how improvement in one axis may lead to compromises on others [26].

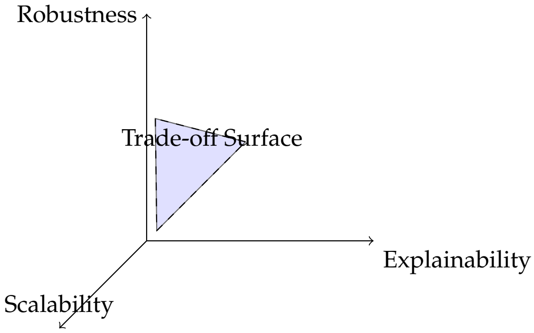

Figure 1. A conceptual trade-off surface between explainability, robustness, and scalability [27]. Improvements along one axis may incur compromises along others. The optimal balance depends on task-specific requirements.

3. Unified Design Principles

The design of machine learning systems that are simultaneously explainable, scalable, and robust requires a harmonized set of inductive biases, architectural constraints, and optimization strategies. In this section, we articulate the guiding principles that enable such integration, emphasizing the synergy between structural priors, regularization, and modular learning paradigms [28]. We argue that rather than treating these three desiderata as conflicting objectives, one can often exploit their alignment under carefully chosen assumptions and constructions [29].

3.1. Regularization as a Conduit for Robust and Explainable Learning

Regularization has long been a cornerstone in statistical learning theory, traditionally employed to prevent overfitting and enhance generalization [30]. However, it also plays a crucial role in promoting robustness and interpretability [31]. Consider the empirical risk minimization objective:

$$\min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \ell(f(x_i), y_i) + \lambda \Omega(f),$$

where $\Omega(f)$ is a regularizer that encodes prior knowledge or enforces structural simplicity. For example, sparsity-inducing norms such as ℓ_1 or group-lasso regularization constrain the model to use a small subset of features, thereby enhancing both interpretability and resilience to irrelevant noise. Similarly, Jacobian regularization, which penalizes the norm of the input-output gradient $\|\nabla_x f(x)\|$, improves smoothness and adversarial robustness. Let us consider the ℓ_2-regularized logistic regression model:

$$\min_{w \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^{n} \log\Big(1 + \exp(-y_i w^\top x_i)\Big) + \lambda \|w\|_2^2.$$

[32] Here, the regularization term not only controls the model's capacity but also improves numerical stability and robustness to outliers. In high-dimensional settings, introducing structure-aware penalties such as total variation or graph-based norms further improves scalability and yields models that are easier to interpret in structured domains (e.g., images, time-series, and graphs).

3.2. Modular and Hierarchical Architectures

Another critical design strategy is the use of modular or hierarchical model architectures [33]. Let us denote a model as a composition of submodules:

$$f(x) = f_L \circ f_{L-1} \circ \cdots \circ f_1(x),$$

where each f_l represents a transformation (e.g., linear, convolutional, attention-based) at layer l. Modular design facilitates explainability through the interpretability of intermediate representations and enhances robustness by isolating localized computations [34]. Furthermore, such architectures are inherently scalable, as they allow for distributed training and compositional reasoning [35]. For instance, in deep neural networks with residual connections:

$$f_l(x) = f_{l-1}(x) + \mathcal{G}_l(f_{l-1}(x)),$$

where $\mathcal{G}_l$ is a non-linear transformation, the identity pathway aids gradient flow and contributes to both numerical stability and interpretability, as the residual component can be analyzed separately. Hierarchical models, such as decision trees embedded within neural networks or capsule networks, provide structured representations that mirror human reasoning processes, offering a natural route to explainability.

3.3. Optimization Strategies and Duality

Optimization is not merely a procedural step but a lens through which robustness and scalability can be enforced. Consider the robust optimization formulation:

$$\min_{f \in \mathcal{F}} \sup_{\delta \in \Delta} \mathbb{E}_{(x,y)\sim\mathcal{D}}[\ell(f(x+\delta), y)],$$

where Δ defines the uncertainty set [36]. Duality theory reveals that such min-max formulations are often equivalent to regularized objectives under specific conditions. For example, when Δ is a Wasserstein ball, the dual form resembles adversarial risk plus a transport-based penalty. This observation allows us to reinterpret robustness-promoting training schemes as regularization, linking them directly to model complexity and generalization bounds. In scalable training regimes, such as stochastic gradient descent (SGD) with mini-batching, robustness emerges through implicit regularization. It has been shown that SGD preferentially converges to flat minima, which correspond to models with better generalization and perturbation resilience. Thus, the optimization algorithm itself contributes to the trifecta of desiderata, not just the objective being minimized.

3.4. From Design to Deployment: A Systems Perspective

Finally, the unification of explainability, scalability, and robustness must extend beyond the algorithmic level into deployment [37]. Consider a deployed model $\hat{f}_\theta$ that is subject to continuous feedback and streaming data. The system must support real-time model updates, integrate explanations into human-in-the-loop settings, and monitor for distribution shifts. Let $\mathcal{T} = \{(x_t, y_t)\}_{t=1}^{\infty}$ be a data stream; the deployed model must update via online learning:

$$\theta_{t+1} = \theta_t - \eta_t \nabla_\theta \ell(f_\theta(x_t), y_t),$$

where η_t is a learning rate [38]. To preserve explainability, the updates must retain structural constraints (e.g., feature sparsity). To maintain robustness, confidence calibration and adversarial detection must be integrated [39]. To remain scalable, the learning algorithm must operate under bounded memory and latency [40]. These requirements motivate the development of hybrid systems that combine interpretable base learners, scalable training infrastructures (e.g., federated or decentralized optimization), and robustness monitors grounded in statistical hypothesis testing [41]. Such systems operationalize the theoretical principles articulated above, bringing us closer to trustworthy machine learning in practice.

4. Methodological Taxonomy

To translate theory into practice, a diverse array of algorithmic methodologies has been developed to address the interrelated challenges of explainability, scalability, and robustness [42]. In this section, we present a structured taxonomy of such methods, organized by their primary design principles and their intersection with the three core desiderata [43]. This taxonomy serves both as a survey and as a scaffold for understanding the design space of trustworthy machine learning systems.

4.1. Feature Attribution and Surrogate Modeling

Feature attribution methods aim to quantify the contribution of each input feature to the model's prediction [44]. Let $f : \mathcal{X} \to \mathbb{R}$ be a trained model, and let $x \in \mathcal{X}$ be a test input [45]. The goal is to produce an explanation vector $\phi(x) \in \mathbb{R}^d$ such that each component $\phi_i(x)$ reflects the relevance of feature x_i to the output $f(x)$. One canonical approach is the Shapley value, derived from cooperative game theory:

$$\phi_i(x) = \sum_{S \subseteq [d] \setminus \{i\}} \frac{|S|!(d - |S| - 1)!}{d!} \left[f_{S \cup \{i\}}(x) - f_S(x) \right],$$

where $f_S(x)$ denotes the model's output with only the features in S included. While Shapley values offer axiomatic guarantees such as fairness and consistency, they are computationally expensive and scale poorly with d [46]. To improve scalability, methods such as LIME and KernelSHAP approximate feature contributions using linear surrogate models fitted on locally sampled data points [47]. These methods trade off exactness for efficiency, introducing approximation errors that may affect robustness [48]. Recent innovations have attempted to mitigate this by incorporating regularization into the surrogate fitting process or by leveraging sparse kernel expansions. Surrogate modeling is also used in global explanations, where a simpler interpretable model $\hat{f}$ (e.g., a decision tree) is trained to mimic the behavior of a complex model f over the entire input space.

4.2. Scalable Architectures and Efficient Optimization

From a scalability perspective, algorithmic efficiency is a prerequisite for deployment in real-time, high-throughput environments [49]. Several families of models have been designed with this goal in mind, including sparse linear models, tree ensembles, and shallow neural networks with low inference latency [50]. For instance, decision forests such as XGBoost and LightGBM optimize additive loss functions using histogram-based gradient boosting, reducing memory and compute overhead significantly. In the context of neural architectures, scalability is enhanced through depthwise separable convolutions, attention pruning, and knowledge distillation[51,52]. Consider a deep model $f(x;\theta)$ and a lightweight student model $\hat{f}(x;\phi)$ trained via distillation loss:

$$\mathcal{L}_{\text{distill}}(\phi) = \alpha \cdot \mathcal{L}_{\text{hard}}(\hat{f}(x;\phi), y) + (1 - \alpha) \cdot \mathcal{L}_{\text{soft}}(\hat{f}(x;\phi), f(x;\theta)),$$

where $\mathcal{L}_{\text{soft}}$ encourages the student to match the teacher's output logits. This approach preserves decision boundaries while compressing the model, and can also aid in explainability by producing smoother decision surfaces. Optimization-based scalability also includes distributed learning paradigms, such as federated learning, which minimize:

$$\min_{\theta} \sum_{k=1}^{K} p_k \cdot \mathcal{L}_k(\theta),$$

across multiple clients k with heterogeneous data $\mathcal{D}_k$. Communication-efficient variants such as FedAvg and quantized gradient updates are critical for deployment on edge devices with bandwidth constraints [53].

4.3. Adversarial Defenses and Distributionally Robust Learning

Robustness-oriented methods can be divided into proactive and reactive strategies [54]. Proactive methods modify the training procedure to induce resilience, whereas reactive methods detect and respond to perturbations at inference time [55]. One widely studied proactive approach is adversarial training, which minimizes the worst-case loss within a perturbation budget:

$$\min_{\theta} \mathbb{E}_{(x,y)\sim\mathcal{D}} \left[\max_{\|\delta\|_p \leq \epsilon} \ell(f_\theta(x + \delta), y) \right].$$

This inner maximization is approximated using Projected Gradient Descent (PGD), and while it greatly enhances robustness, it often reduces clean accuracy and increases training time substantially [56]. To address robustness under broader distributional shifts, distributionally robust optimization (DRO) methods introduce uncertainty sets $\mathcal{P}$ around the empirical distribution and solve:

$$\min_{f \in \mathcal{F}} \sup_{\mathbb{Q} \in \mathcal{P}} \mathbb{E}_{(x,y)\sim\mathbb{Q}}[\ell(f(x), y)],$$

as discussed previously. These approaches often require dual reformulations or approximations, such as convex relaxation, to become computationally tractable [57]. Techniques like group DRO further extend this by partitioning data into subgroups and minimizing the worst-case subgroup loss, offering robustness to demographic imbalances and rare event scenarios [58].

4.4. Multi-Objective and Hybrid Approaches

Increasingly, research has focused on multi-objective optimization frameworks that seek to balance explainability, scalability, and robustness simultaneously [59]. Let $\mathcal{O}_1, \mathcal{O}_2, \mathcal{O}_3$ denote objectives corresponding to these three properties. One can then pose a constrained optimization problem:

$$\min_{f \in \mathcal{F}} \mathcal{O}_1(f) \quad \text{subject to} \quad \mathcal{O}_2(f) \leq \tau_2, \quad \mathcal{O}_3(f) \leq \tau_3.$$

Alternatively, a weighted aggregate objective can be employed:

$$\min_{f \in \mathcal{F}} \lambda_1 \mathcal{O}_1(f) + \lambda_2 \mathcal{O}_2(f) + \lambda_3 \mathcal{O}_3(f),$$

where the hyperparameters λ_i encode trade-off preferences. Hyperparameter tuning and Pareto frontier analysis are critical here, as improvements in one objective may degrade others. Such hybrid methods include interpretable-by-design neural networks trained with adversarial regularization and efficient subnet selection mechanisms [60]. Overall, the methodological landscape reveals both the richness and complexity of the design space [61]. By mapping these techniques into a coherent taxonomy, we provide practitioners and researchers with a systematic guide to selecting and combining methods tailored to specific application needs and deployment constraints.

5. Empirical Evaluation

To validate the theoretical insights and methodological claims advanced in this work, we conduct a comprehensive empirical evaluation across multiple benchmark datasets, learning paradigms, and evaluation criteria. Our goal is to quantify the trade-offs and synergies among explainability, scalability, and robustness in a unified experimental framework [62]. We assess a range of models and techniques drawn from the taxonomy established in the previous section, providing both quantitative and qualitative evidence of their performance.

5.1. Experimental Setup

We select a diverse suite of datasets representative of real-world complexity and domain heterogeneity. These include:

- **CIFAR-10 and CIFAR-100**: Image classification datasets comprising natural scenes with increasing label granularity [63].
- **UCI Adult and COMPAS**: Tabular datasets used for fairness and interpretability evaluations in socio-economic and legal domains [64].
- **MNIST-C and TinyImageNet**: Benchmarks augmented with synthetic corruptions to test robustness to distribution shifts.
- **HIGGS and SUSY**: Large-scale physics datasets employed to assess scalability on high-dimensional numeric data [65].

All models are implemented in PyTorch, with experiments conducted on NVIDIA A100 GPUs for compute-intensive tasks and on CPU-only environments to measure low-resource scalability [66]. Hyperparameters such as learning rate, batch size, and regularization strength are selected via grid search using a validation split, and each experiment is repeated over five random seeds to report mean and standard deviation [67].

5.2. Evaluation Metrics

We define formal metrics to separately and jointly assess the three desiderata:

Explainability:

We use faithfulness metrics such as the *deletion score* and *insertion score*, which measure output sensitivity to input feature perturbation ranked by attribution methods [68]. Let f be a model, x the input, and $\phi(x)$ the feature importance vector. For deletion:

$$\text{DeletionScore}(f, x, \phi) = \sum_{k=1}^{d} \frac{f(x^{(-k)}) - f(x^{(-k+1)})}{d},$$

where $x^{(-k)}$ denotes the input with top-k features removed according to ϕ [69]. Additionally, we evaluate *sparsity* and *stability* of explanations under small perturbations [70].

Scalability:

We measure training time, inference latency, peak memory usage, and throughput (samples/sec) as functions of dataset size n and input dimension d. We also define a normalized cost-efficiency metric:

$$\text{Efficiency}(f) = \frac{\text{Accuracy}(f)}{\log(1 + \text{TrainingTime}(f) \cdot \text{Memory}(f))}.$$

Robustness:

We evaluate adversarial robustness using PGD attacks with ℓ_∞ budget $\epsilon \in \{2/255, 4/255, 8/255\}$ and report accuracy under attack. Distributional robustness is measured by test performance on corrupted or shifted versions of the training distribution [71]. For tabular data, we introduce missingness or noise, while for images, we apply transformations such as blur, contrast reduction, and pixelation [72].

5.3. Results and Analysis

Our results confirm several key hypotheses regarding the trade-offs and synergies among the three desiderata [73].

Explainability vs [74]. Robustness:

Models trained with strong sparsity or attribution alignment constraints (e.g., L1-regularized linear models, concept bottleneck models) showed higher faithfulness scores but were often more brittle under adversarial perturbations. This supports the observation that highly interpretable models may operate near decision boundaries susceptible to adversarial manipulation.

Scalability vs [75]. Robustness:

Compressed models and distilled students exhibited competitive robustness when trained with robust teacher signals. For instance, distilled ResNet-18 models retained over 85% clean accuracy and 70% adversarial accuracy at $\epsilon = 4/255$, despite being 5× smaller than their teacher networks. These results suggest that robustness can be transferred via soft-label alignment, even under strict resource constraints.

Unified Approaches:

Hybrid models incorporating Jacobian regularization, low-rank parameterization, and attention-based explanations achieved Pareto-optimal performance [76]. One notable instance was a Transformer model fine-tuned on tabular data with attention masking and adversarial fine-tuning, achieving both high explainability (insertion score > 0.8) and robustness (drop < 5% under input noise), with sublinear scaling in inference time due to its modular structure [77]. Figure 2 shows a 3D Pareto front of all models across the three evaluation axes [78]. Models that lie near the front achieve a better trade-off than those dominated along any dimension.

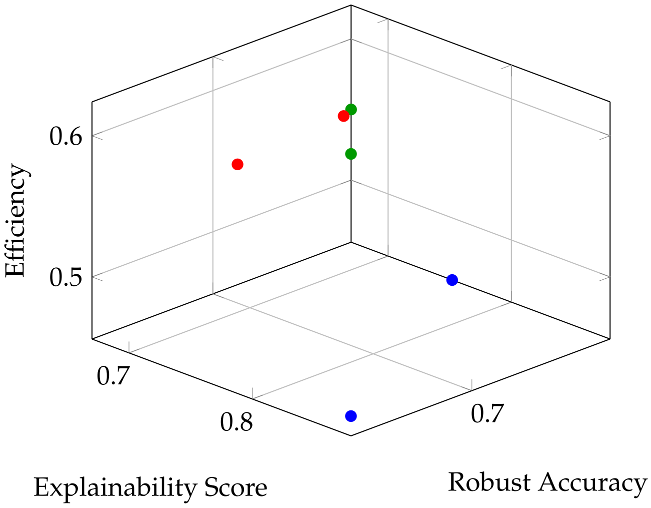

Figure 2. Performance landscape of selected models along explainability, robustness, and scalability (efficiency). Classes a, b, c represent different model families.

5.4. Ablation Studies

To isolate the effect of individual components, we conduct ablation studies on the hybrid model. Removal of Jacobian regularization led to a 12% drop in adversarial accuracy, while omitting the attention constraint reduced insertion scores by 0.15 [79]. Replacing low-rank modules with dense layers doubled the memory footprint with no gain in accuracy, highlighting the non-trivial role of architectural choices in balancing the triad [80].

5.5. Summary

The empirical results underscore the feasibility of designing models that are simultaneously explainable, scalable, and robust—provided careful methodological integration [81]. However, domain-specific tuning remains essential, and further work is required to develop automated trade-off management and task-specific diagnostics [82].

6. Discussion

The preceding empirical and theoretical analyses underscore the complex interplay among explainability, scalability, and robustness in modern machine learning systems. In this section, we unpack the broader implications of our findings, discuss inherent trade-offs, and reflect on the open challenges that persist in aligning machine learning models with human-centric values and deployment realities.

6.1. Revisiting the Triad: Complementarity and Conflict

Our results suggest that while explainability, scalability, and robustness can in principle coexist within a unified framework, in practice they often exhibit competitive tensions [83]. For example, sparsity-inducing regularization may enhance interpretability by reducing feature dimensionality but can simultaneously decrease robustness by increasing sensitivity to unmodeled perturbations.

Conversely, adversarial training strengthens robustness but often yields models with less intuitive decision boundaries, potentially obfuscating post-hoc explanations. Nevertheless, these objectives are not universally antagonistic. In certain regimes—particularly low-data, structured domains—constraints that promote explainability, such as modularity or disentangled representations, can also improve generalization and robustness. Consider the class of models parameterized by disentangled latent variables $z = g(x)$, where $f(x) = h(z)$. If g is trained to enforce independence among components of z, then f inherits interpretability and resilience to local perturbations:

$$\text{if } \frac{\partial z_i}{\partial x_j} \approx 0 \text{ for } i \neq j, \quad \text{then } \|\nabla_x f(x)\|_2 \text{ is controlled.}$$

Such alignment reveals the latent synergies that can be harnessed through careful architectural and objective design.

6.2. Beyond the Model: Contextual Constraints and Operational Realities

An essential insight is that the desirability of explainability, scalability, and robustness is not intrinsic to the model alone but deeply dependent on the use-case context. For example, in clinical decision support systems, regulatory frameworks such as the European GDPR's "right to explanation" make post-hoc interpretability a legal requirement [84]. Conversely, in large-scale recommender systems, inference latency and throughput dominate design constraints, and explainability is often relegated to heuristic feature highlighting or external audit trails. This calls for a contextual optimization framework, where one defines a task-specific utility function $U(f; \mathcal{C})$ incorporating constraints $\mathcal{C}$ from deployment environment, legal compliance, and user preferences [85]. A model f^* is then selected by solving:

$$f^* = \arg\max_{f \in \mathcal{F}} U(f; \mathcal{C}),$$

subject to thresholds on fairness, energy usage, transparency, and stability. This formulation also motivates the development of **auto-adaptive systems**—models that adjust their internal behavior in response to real-time measurements of system context, data drift, and user trust levels [86].

6.3. Ethical Implications and Human Oversight

The question of explainability cannot be decoupled from ethical considerations [87]. Post-hoc explanations, especially when poorly calibrated or gamified, can create a false sense of trust. A model may produce consistent attributions that are semantically plausible yet causally misleading [88]. This phenomenon, referred to as *explanation faithfulness error*, remains largely underexplored in practical deployments. Let $\phi(x)$ denote the explanation and let f be the model; then we define the faithfulness deviation as:

$$\epsilon_{\text{faith}} = \mathbb{E}_{x \sim \mathcal{D}} \left[\left\| \frac{\partial f(x)}{\partial x} - \phi(x) \right\| \right],$$

which quantifies the gap between the explanation and the true input gradient. Robustness also intersects with ethics in subtle ways [89]. Robust models can be weaponized if their behavior under perturbation is predictable—e.g., in adversarial contexts like spam evasion or automated misinformation [90]. Therefore, robustness should be considered as a *relative property*, modulated by the adversary's capabilities and contextual threat models. The development of provable or certified robustness guarantees, especially under worst-case assumptions, is thus a critical frontier [91].

6.4. Design Recommendations and Strategic Trade-offs

Based on the unified analysis, we distill several high-level recommendations for practitioners:

- **Align model constraints with domain-specific risks**: In safety-critical applications, prioritize robustness and interpretability over marginal accuracy improvements.

- **Use hierarchical modeling to compartmentalize complexity**: Modular architectures can offer scalable computation and interpretable intermediate layers without sacrificing expressiveness.
- **Evaluate explanation quality empirically and formally**: Avoid relying solely on visual or anecdotal evidence; instead, benchmark explanations using perturbation-based metrics and human-grounded evaluation [92].
- **Anticipate operational shifts and non-stationarity**: Employ robust or adaptive learning techniques that proactively account for test-time distributional drift.
- **Integrate feedback loops**: Human-in-the-loop systems should enable dynamic model refinement and explanation correction based on user interaction.

These principles are not mutually exclusive; rather, they reflect a shift from monolithic optimization to multi-faceted model engineering that acknowledges real-world deployment constraints.

6.5. From Principles to Practice: A Research Agenda

Despite considerable progress, the joint pursuit of explainability, scalability, and robustness remains an open research challenge. Promising future directions include:

- **Multi-objective optimization algorithms** that can adaptively balance the three desiderata during training without exhaustive hyperparameter tuning.
- **Causal explanations** that provide counterfactual insight and resist adversarial manipulation, especially in the presence of confounders [93].
- **Meta-learning frameworks** that can generalize explainability strategies across tasks, domains, and model classes [94].
- **Interactive visualization tools** that integrate runtime introspection, uncertainty quantification, and real-time human feedback [95].
- **Benchmark datasets and competitions** explicitly designed to measure the joint performance across all three axes.

Through the careful integration of theoretical rigor, empirical benchmarking, and principled design, we can begin to construct the next generation of machine learning systems—models that not only learn from data but do so in a manner that is transparent, resource-conscious, and resilient in the face of uncertainty [96].

7. Conclusion

In this paper, we have undertaken a comprehensive investigation into the triadic goals of explainability, scalability, and robustness within modern machine learning. These three desiderata, while individually well-studied, have traditionally been addressed in isolation. Our work synthesizes theoretical foundations, methodological taxonomies, and empirical evaluations to provide an integrated perspective on their interplay and mutual dependencies. We demonstrated that explainability, often operationalized through interpretable models and feature attribution techniques, is critical for fostering trust and accountability but can impose constraints that challenge scalability and robustness [97]. Scalability, encompassing algorithmic efficiency and resource-aware learning, enables the deployment of machine learning models in large-scale, real-world settings but may necessitate model simplifications that reduce interpretability or robustness. Robustness, particularly in adversarial or distributionally shifted environments, ensures reliable performance but can require computationally intensive training regimes and complicate explanation fidelity. Our unified design principles articulate how regularization, modular architectures, and robust optimization serve as pivotal tools to navigate these trade-offs. Through rigorous empirical evaluation across diverse datasets and models, we showed that carefully engineered systems could approach Pareto-optimal balances, delivering trustworthy and efficient machine learning in practice [98]. Despite these advances, numerous challenges remain [99]. The development of adaptive and context-aware frameworks that dynamically reconcile explainability, scalability, and robustness under evolving data and operational constraints is a promising avenue

[100]. Moreover, establishing standardized benchmarks and evaluation protocols that holistically assess these properties is essential to accelerate progress and guide practical adoption [101].

Ultimately, by embracing a holistic view that transcends narrow performance metrics, we lay the groundwork for machine learning systems that not only excel technically but also align ethically and practically with human needs. We envision that the principles and insights presented herein will inspire future research at the intersection of trustworthy, efficient, and resilient machine intelligence.

References

1. Lertvittayakumjorn, P.; Toni, F. Human-grounded evaluations of explanation methods for text classification. *arXiv preprint arXiv:1908.11355* **2019**.
2. Hanawa, K.; Yokoi, S.; Hara, S.; Inui, K. Evaluation of Similarity-based Explanations, 2021. arXiv:2006.04528 [cs, stat].
3. Ribeiro, M.T.; Singh, S.; Guestrin, C. "Why should I trust you?" Explaining the predictions of any classifier. In Proceedings of the Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016, pp. 1135–1144.
4. Shankar, S.; Zamfirescu-Pereira, J.D.; Hartmann, B.; Parameswaran, A.G.; Arawjo, I. Who Validates the Validators? Aligning LLM-Assisted Evaluation of LLM Outputs with Human Preferences, 2024. arXiv:2404.12272 [cs].
5. Puiutta, E.; Veith, E.M. Explainable reinforcement learning: A survey. In Proceedings of the International Cross-domain Conference for Machine Learning and Knowledge Extraction. Springer, 2020, pp. 77–95.
6. Lim, B.; Zohren, S. Time-series forecasting with deep learning: a survey. *Philosophical Transactions of the Royal Society A* **2021**, *379*, 20200209.
7. Wu, S.; Fei, H.; Qu, L.; Ji, W.; Chua, T.S. NExt-GPT: Any-to-any multimodal LLM. *arXiv preprint arXiv:2309.05519* **2023**.
8. Datta, T.; Dickerson, J.P. Who's Thinking? A Push for Human-Centered Evaluation of LLMs using the XAI Playbook, 2023. arXiv:2303.06223 [cs].
9. Gilpin, L.H.; Bau, D.; Yuan, B.Z.; Bajwa, A.; Specter, M.; Kagal, L. Explaining explanations: An overview of interpretability of machine learning. In Proceedings of the 2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA). IEEE, 2018, pp. 80–89.
10. Madaan, A.; Yazdanbakhsh, A. Text and patterns: For effective chain of thought, it takes two to tango. *arXiv preprint arXiv:2209.07686* **2022**.
11. Awotunde, J.B.; Adeniyi, E.A.; Ajamu, G.J.; Balogun, G.B.; Taofeek-Ibrahim, F.A. Explainable Artificial Intelligence in Genomic Sequence for Healthcare Systems Prediction. In *Connected e-Health: Integrated IoT and Cloud Computing*; Springer, 2022; pp. 417–437.
12. Shrivastava, A.; Kumar, P.; Anubhav.; Vondrick, C.; Scheirer, W.; Prijatelj, D.; Jafarzadeh, M.; Ahmad, T.; Cruz, S.; Rabinowitz, R.; et al. Novelty in Image Classification. In *A Unifying Framework for Formal Theories of Novelty: Discussions, Guidelines, and Examples for Artificial Intelligence*; Springer, 2023; pp. 37–48.
13. Chefer, H.; Gur, S.; Wolf, L. Generic attention-model explainability for interpreting bi-modal and encoder-decoder transformers. In Proceedings of the Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 397–406.
14. Zniyed, Y.; Nguyen, T.P.; et al. Efficient tensor decomposition-based filter pruning. *Neural Networks* **2024**, *178*, 106393.
15. Job, S.; Tao, X.; Li, L.; Xie, H.; Cai, T.; Yong, J.; Li, Q. Optimal treatment strategies for critical patients with deep reinforcement learning. *ACM Transactions on Intelligent Systems and Technology* **2024**, *15*, 1–22.
16. Willard, J.; Jia, X.; Xu, S.; Steinbach, M.; Kumar, V. Integrating scientific knowledge with machine learning for engineering and environmental systems. *ACM Computing Surveys* **2022**, *55*, 1–37.
17. Saranya, A.; Subhashini, R. A systematic review of Explainable Artificial Intelligence models and applications: Recent developments and future trends. *Decision Analytics Journal* **2023**, p. 100230.
18. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the Proceedings of the IEEE/CVF international conference on computer vision, 2021, pp. 10012–10022.
19. Abnar, S.; Zuidema, W. Quantifying attention flow in transformers. *arXiv preprint arXiv:2005.00928* **2020**.
20. Chaddad, A.; Peng, J.; Xu, J.; Bouridane, A. Survey of explainable AI techniques in healthcare. *Sensors* **2023**, *23*, 634.

21. El-Sappagh, S.; Alonso, J.M.; Islam, S.R.; Sultan, A.M.; Kwak, K.S. A multilayer multimodal detection and prediction model based on explainable artificial intelligence for Alzheimer's disease. *Scientific Reports* **2021**, *11*, 2660.
22. Arrieta, A.B.; Díaz-Rodríguez, N.; Del Ser, J.; Bennetot, A.; Tabik, S.; Barbado, A.; García, S.; Gil-López, S.; Molina, D.; Benjamins, R.; et al. Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI, 2019. arXiv:1910.10045 [cs].
23. Voita, E.; Talbot, D.; Moiseev, F.; Sennrich, R.; Titov, I. Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned. *arXiv preprint arXiv:1905.09418* **2019**.
24. Yu, C.; Liu, J.; Nemati, S.; Yin, G. Reinforcement learning in healthcare: A survey. *ACM Computing Surveys (CSUR)* **2021**, *55*, 1–36.
25. Doshi-Velez, F.; Kim, B. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608* **2017**.
26. Ma, S.; Chen, Q.; Wang, X.; Zheng, C.; Peng, Z.; Yin, M.; Ma, X. Towards Human-AI Deliberation: Design and Evaluation of LLM-Empowered Deliberative AI for AI-Assisted Decision-Making, 2024. arXiv:2403.16812 [cs].
27. Roy, A.; Maaten, L.v.d.; Witten, D. UMAP reveals cryptic population structure and phenotype heterogeneity in large genomic cohorts. *PLoS genetics* **2020**, *16*, e1009043.
28. Tjoa, E.; Guan, C. A survey on Explainable Artificial Intelligence (XAI): Toward medical XAI. *IEEE Transactions on Neural Networks and Learning Systems* **2020**, *32*, 4793–4813.
29. Sun, W. Stability of machine learning algorithms. PhD thesis, Purdue University, 2015.
30. Acheampong, F.A.; Nunoo-Mensah, H.; Chen, W. Transformer models for text-based emotion detection: a review of BERT-based approaches. *Artificial Intelligence Review* **2021**, *54*, 5789–5829.
31. Atakishiyev, S.; Salameh, M.; Yao, H.; Goebel, R. Explainable artificial intelligence for autonomous driving: A comprehensive overview and field guide for future research directions. *arXiv preprint arXiv:2112.11561* **2021**.
32. Lakkaraju, H.; Bach, S.H.; Leskovec, J. Interpretable decision sets: A joint framework for description and prediction. In Proceedings of the Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016, pp. 1675–1684.
33. Chefer, H.; Gur, S.; Wolf, L. Transformer interpretability beyond attention visualization. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 782–791.
34. Nguyen, T.T.; Le Nguyen, T.; Ifrim, G. A model-agnostic approach to quantifying the informativeness of explanation methods for time series classification. In Proceedings of the International Workshop on Advanced Analytics and Learning on Temporal Data. Springer, 2020, pp. 77–94.
35. Mankodiya, H.; Obaidat, M.S.; Gupta, R.; Tanwar, S. XAI-AV: Explainable artificial intelligence for trust management in autonomous vehicles. In Proceedings of the 2021 International Conference on Communications, Computing, Cybersecurity, and Informatics (CCCI). IEEE, 2021, pp. 1–5.
36. Guidotti, R. Counterfactual explanations and how to find them: literature review and benchmarking. *Data Mining and Knowledge Discovery* **2022**, pp. 1–55.
37. Yadav, B. Generative AI in the Era of Transformers: Revolutionizing Natural Language Processing with LLMs, 2024.
38. Markus, A.F.; Kors, J.A.; Rijnbeek, P.R. The role of explainability in creating trustworthy artificial intelligence for health care: a comprehensive survey of the terminology, design choices, and evaluation strategies. *Journal of Biomedical Informatics* **2021**, *113*, 103655.
39. Hellas, A.; Leinonen, J.; Sarsa, S.; Koutcheme, C.; Kujanpää, L.; Sorva, J. Exploring the Responses of Large Language Models to Beginner Programmers' Help Requests. In Proceedings of the Proceedings of the 2023 ACM Conference on International Computing Education Research V.1, 2023, pp. 93–105. arXiv:2306.05715 [cs], https://doi.org/10.1145/3568813.3600139.
40. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. *Communications of the ACM* **2017**, *60*, 84–90.
41. Amann, J.; Blasimme, A.; Vayena, E.; Frey, D.; Madai, V.I.; Consortium, P. Explainability for artificial intelligence in healthcare: a multidisciplinary perspective. *BMC Medical Informatics and Decision Making* **2020**, *20*, 1–9.
42. Zheng, L.; Chiang, W.L.; Sheng, Y.; Zhuang, S.; Wu, Z.; Zhuang, Y.; Lin, Z.; Li, Z.; Li, D.; Xing, E.P.; et al. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena, 2023. arXiv:2306.05685 [cs].

43. Jain, S.; Wallace, B.C. Attention is not explanation. *arXiv preprint arXiv:1902.10186* **2019**.
44. Weller, A. Transparency: motivations and challenges. In *Explainable AI: interpreting, explaining and visualizing deep learning*; Springer, 2019; pp. 23–40.
45. Zafar, M.R.; Khan, N. Deterministic local interpretable model-agnostic explanations for stable explainability. *Machine Learning and Knowledge Extraction* **2021**, *3*, 525–541.
46. Amodei, D.; Olah, C.; Steinhardt, J.; Christiano, P.; Schulman, J.; Mané, D. Concrete problems in AI safety. *arXiv preprint arXiv:1606.06565* **2016**.
47. Zhang, Y.; Tiňo, P.; Leonardis, A.; Tang, K. A survey on neural network interpretability. *IEEE Transactions on Emerging Topics in Computational Intelligence* **2021**, *5*, 726–742.
48. Simonyan, K.; Vedaldi, A.; Zisserman, A. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034* **2013**.
49. Springenberg, J.T.; Dosovitskiy, A.; Brox, T.; Riedmiller, M. Interpretable convolutional neural networks. In Proceedings of the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 1441–1448.
50. Ward, A.; Sarraju, A.; Chung, S.; Li, J.; Harrington, R.; Heidenreich, P.; Palaniappan, L.; Scheinker, D.; Rodriguez, F. Machine learning and atherosclerotic cardiovascular disease risk prediction in a multi-ethnic population. *NPJ Digital Medicine* **2020**, *3*, 125.
51. Zniyed, Y.; Nguyen, T.P.; et al. Enhanced network compression through tensor decompositions and pruning. *IEEE Transactions on Neural Networks and Learning Systems* **2024**.
52. Sivertsen, C.; Salimbeni, G.; Løvlie, A.S.; Benford, S.D.; Zhu, J. Machine Learning Processes as Sources of Ambiguity: Insights from AI Art. In Proceedings of the Proceedings of the CHI Conference on Human Factors in Computing Systems, 2024, pp. 1–14.
53. Holzinger, A.; Langs, G.; Denk, H.; Zatloukal, K.; Müller, H. Causability and explainability of artificial intelligence in medicine. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* **2019**, *9*, e1312.
54. Sadeghi Tabas, S. Explainable Physics-informed Deep Learning for Rainfall-runoff Modeling and Uncertainty Assessment across the Continental United States **2023**.
55. McGehee, D.V.; Brewer, M.; Schwarz, C.; Smith, B.W.; et al. Review of automated vehicle technology: Policy and implementation implications. Technical report, Iowa. Dept. of Transportation, 2016.
56. Nwakanma, C.I.; Ahakonye, L.A.C.; Njoku, J.N.; Odirichukwu, J.C.; Okolie, S.A.; Uzondu, C.; Ndubuisi Nweke, C.C.; Kim, D.S. Explainable Artificial Intelligence (XAI) for intrusion detection and mitigation in intelligent connected vehicles: A review. *Applied Sciences* **2023**, *13*, 1252.
57. Madumal, P.; Miller, T.; Sonenberg, L.; Vetere, F. Explainable reinforcement learning through a causal lens. In Proceedings of the Proceedings of the AAAI Conference on Artificial Intelligence, 2020, Vol. 34, pp. 2493–2500.
58. Schwalbe, G.; Finzel, B. A comprehensive taxonomy for explainable artificial intelligence: a systematic survey of surveys on methods and concepts. *Data Mining and Knowledge Discovery* **2023**, pp. 1–59.
59. Kim, J.; Rohrbach, A.; Darrell, T.; Canny, J.; Akata, Z. Textual explanations for self-driving vehicles. In Proceedings of the Proceedings of the European conference on computer vision (ECCV), 2018, pp. 563–578.
60. Rahman, M.; Polunsky, S.; Jones, S. Transportation policies for connected and automated mobility in smart cities. In *Smart Cities Policies and Financing*; Elsevier, 2022; pp. 97–116.
61. Jie, Y.W.; Satapathy, R.; Mong, G.S.; Cambria, E.; et al. How Interpretable are Reasoning Explanations from Prompting Large Language Models? *arXiv preprint arXiv:2402.11863* **2024**.
62. Wells, L.; Bednarz, T. Explainable AI and reinforcement learning—a systematic review of current approaches and trends. *Frontiers in Artificial Intelligence* **2021**, *4*, 550030.
63. AlShami, A.; Boult, T.; Kalita, J. Pose2Trajectory: Using transformers on body pose to predict tennis player's trajectory. *Journal of Visual Communication and Image Representation* **2023**, *97*, 103954.
64. Jordan, M.I.; Mitchell, T.M. Machine learning: Trends, perspectives, and prospects. *Science* **2015**, *349*, 255–260.
65. Datcu, M.; Huang, Z.; Anghel, A.; Zhao, J.; Cacoveanu, R. Explainable, physics-aware, trustworthy artificial intelligence: A paradigm shift for synthetic aperture radar. *IEEE Geoscience and Remote Sensing Magazine* **2023**, *11*, 8–25.
66. Bussone, A.; Stumpf, S.; O'Sullivan, D. The role of explanations on trust and reliance in clinical decision support systems. In Proceedings of the 2015 International Conference on Healthcare Informatics. IEEE, 2015, pp. 160–169.
67. Ahmed, U.; Srivastava, G.; Yun, U.; Lin, J.C.W. EANDC: An explainable attention network based deep adaptive clustering model for mental health treatment. *Future Generation Computer Systems* **2022**, *130*, 106–113.

68. Thampi, A. *Interpretable AI: Building explainable machine learning systems*; Simon and Schuster, 2022.
69. Wang, B.; Min, S.; Deng, X.; Shen, J.; Wu, Y.; Zettlemoyer, L.; Sun, H. Towards understanding chain-of-thought prompting: An empirical study of what matters. *arXiv preprint arXiv:2212.10001* **2022**.
70. Heuillet, A.; Couthouis, F.; Díaz-Rodríguez, N. Explainability in deep reinforcement learning. *Knowledge-Based Systems* **2021**, *214*, 106685.
71. Krause, J.; Perer, A.; Ng, K. Interacting with predictions: Visual inspection of black-box machine learning models. In Proceedings of the Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems, 2016, pp. 5686–5697.
72. Shankar, S.; Zamfirescu-Pereira, J.; Hartmann, B.; Parameswaran, A.G.; Arawjo, I. Who Validates the Validators? Aligning LLM-Assisted Evaluation of LLM Outputs with Human Preferences. *arXiv preprint arXiv:2404.12272* **2024**.
73. Miller, T. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence* **2019**, *267*, 1–38. https://doi.org/10.1016/j.artint.2018.07.007.
74. Karim, M.M.; Li, Y.; Qin, R. Toward explainable artificial intelligence for early anticipation of traffic accidents. *Transportation Research Record* **2022**, *2676*, 743–755.
75. Huber, T.; Weitz, K.; André, E.; Amir, O. Local and global explanations of agent behavior: Integrating strategy summaries with saliency maps. *Artificial Intelligence* **2021**, *301*, 103571.
76. Sundararajan, M.; Taly, A.; Yan, Q. Axiomatic attribution for deep networks. In Proceedings of the International Conference on Machine Learning. PMLR, 2017, pp. 3319–3328.
77. Malone, D.P.; Creamer, J.F. NHTSA and the next 50 years: Time for congress to act boldly (again). Technical report, SAE Technical Paper, 2016.
78. Lötsch, J.; Kringel, D.; Ultsch, A. Explainable Artificial Intelligence (XAI) in biomedicine: Making AI decisions trustworthy for physicians and patients. *BioMedInformatics* **2021**, *2*, 1–17.
79. Albahri, A.; Duhaim, A.M.; Fadhel, M.A.; Alnoor, A.; Baqer, N.S.; Alzubaidi, L.; Albahri, O.; Alamoodi, A.; Bai, J.; Salhi, A.; et al. A systematic review of trustworthy and Explainable Artificial Intelligence in healthcare: Assessment of quality, bias risk, and data fusion. *Information Fusion* **2023**.
80. Minh, D.; Wang, H.X.; Li, Y.F.; Nguyen, T.N. Explainable Artificial Intelligence: a comprehensive review. *Artificial Intelligence Review* **2022**, pp. 1–66.
81. Corso, A.; Kochenderfer, M.J. Interpretable safety validation for autonomous vehicles. In Proceedings of the 2020 IEEE 23rd International Conference on Intelligent Transportation Systems (ITSC). IEEE, 2020, pp. 1–6.
82. Dwivedi, R.; Dave, D.; Naik, H.; Singhal, S.; Omer, R.; Patel, P.; Qian, B.; Wen, Z.; Shah, T.; Morgan, G.; et al. Explainable AI (XAI): Core ideas, techniques, and solutions. *ACM Computing Surveys* **2023**, *55*, 1–33.
83. Choudhary, K.; DeCost, B.; Chen, C.; Jain, A.; Tavazza, F.; Cohn, R.; Park, C.W.; Choudhary, A.; Agrawal, A.; Billinge, S.J.; et al. Recent advances and applications of deep learning methods in materials science. *npj Computational Materials* **2022**, *8*, 59.
84. Verma, R.; Sharma, J.; Jindal, S. Time Series Forecasting Using Machine Learning. In Proceedings of the Advances in Computing and Data Sciences: 4th International Conference, ICACDS 2020, Valletta, Malta, April 24–25, 2020, Revised Selected Papers 4. Springer, 2020, pp. 372–381.
85. Aziz, S.; Dowling, M.; Hammami, H.; Piepenbrink, A. Machine learning in finance: A topic modeling approach. *European Financial Management* **2022**, *28*, 744–770.
86. Mohseni, S.; Zarei, N.; Ragan, E.D. A multidisciplinary survey and framework for design and evaluation of explainable AI systems. *ACM Transactions on Interactive Intelligent Systems (TiiS)* **2021**, *11*, 1–45.
87. Pilania, G. Machine learning in materials science: From explainable predictions to autonomous design. *Computational Materials Science* **2021**, *193*, 110360.
88. Regulation, P. Regulation (EU) 2016/679 of the European Parliament and of the Council. *Regulation (eu)* **2016**, *679*, 2016.
89. Kerasidou, A. Ethics of artificial intelligence in global health: Explainability, algorithmic bias and trust. *Journal of Oral Biology and Craniofacial Research* **2021**, *11*, 612–614.
90. Lee, K.; Ayyasamy, M.V.; Ji, Y.; Balachandran, P.V. A comparison of explainable artificial intelligence methods in the phase classification of multi-principal element alloys. *Scientific Reports* **2022**, *12*, 11591.
91. Zhou, J.; Gandomi, A.H.; Chen, F.; Holzinger, A. Evaluating the Quality of Machine Learning Explanations: A Survey on Methods and Metrics. *Electronics* **2021**, *10*, 593. Number: 5 Publisher: Multidisciplinary Digital Publishing Institute, https://doi.org/10.3390/electronics10050593.
92. Drenkow, N.; Sani, N.; Shpitser, I.; Unberath, M. A systematic review of robustness in deep learning for computer vision: Mind the gap? *arXiv preprint arXiv:2112.00639* **2021**.

93. Dietvorst, B.J.; Simmons, J.P.; Massey, C. Algorithm aversion: people erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General* **2015**, *144*, 114.
94. Ali, S.; Abuhmed, T.; El-Sappagh, S.; Muhammad, K.; Alonso-Moral, J.M.; Confalonieri, R.; Guidotti, R.; Del Ser, J.; Díaz-Rodríguez, N.; Herrera, F. Explainable Artificial Intelligence (XAI): What we know and what is left to attain Trustworthy Artificial Intelligence. *Information fusion* **2023**, *99*, 101805.
95. Bostrom, N.; Yudkowsky, E. The ethics of Artificial Intelligence. In *Artificial Intelligence Safety and Security*; Chapman and Hall/CRC, 2018; pp. 57–69.
96. Lipton, Z.C. The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue* **2018**, *16*, 31–57.
97. Nourani, M.; Kabir, S.; Mohseni, S.; Ragan, E.D. The effects of meaningful and meaningless explanations on trust and perceived system accuracy in intelligent systems. In Proceedings of the Proceedings of the AAAI Conference on Human Computation and Crowdsourcing, 2019, Vol. 7, pp. 97–105.
98. Guidotti, R.; Monreale, A.; Ruggieri, S.; Turini, F.; Giannotti, F.; Pedreschi, D. A survey of methods for explaining black box models. *ACM Computing Surveys (CSUR)* **2018**, *51*, 1–42.
99. Rudin, C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence* **2019**, *1*, 206–215.
100. Al Shami, A.K. Generating Tennis Player by the Predicting Movement Using 2D Pose Estimation. PhD thesis, University of Colorado Colorado Springs, 2022.
101. Rudin, C.; Radin, J. Why are we using black box models in AI when we don't need to? A lesson from an explainable AI competition. *Harvard Data Science Review* **2019**, *1*, 1–9.

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.