Practical Guidelines for Building Explainable, Efficient, and Robust Machine Learning Systems

Huan Zheng[1], Zhihao Ru[1]

Department of Computer Science, Chinese University of Hong Kong, Hong Kong
huan.zheng@cuhk.edu.hk, zhihao.ru@cuhk.edu.hk

Abstract. The growing integration of machine learning (ML) into critical applications has elevated the importance of models that are not only accurate but also *explainable*, *efficient*, and *robust*. While each of these properties has been extensively studied in isolation, their simultaneous realization remains a formidable and largely unresolved challenge. This paper presents a comprehensive exploration of the theoretical, algorithmic, and empirical foundations for constructing ML systems that jointly satisfy these three desiderata. We begin by formalizing the multi-objective learning framework, introducing mathematical formulations that capture the trade-offs among interpretability, computational parsimony, and resilience to perturbations. We then survey a wide spectrum of algorithmic strategies, including sparse modeling, distillation, adversarial training, modular architectures, and multi-objective optimization techniques. Through detailed empirical evaluations across domains such as healthcare, autonomous driving, finance, and natural language processing, we quantify the interdependencies and tensions among the triadic objectives. Our results highlight that no single solution dominates across all metrics, but careful design choices can yield models that approach Pareto-optimal performance in practical settings. Building on these insights, we propose a set of system-level design principles for deploying trustworthy ML, including modularization, continuous monitoring, and human-centered explanation interfaces. We conclude with an agenda for future research, calling for unified theoretical frameworks, domain-aware evaluation protocols, and interdisciplinary collaboration to advance the field toward more transparent, resilient, and accessible AI.

Keywords: Explainable machine learning, interpretable models, model efficiency, computational optimization, robust machine learning, adversarial robustness, multi-objective optimization, model compression, trustworthy AI, human-centered AI, modular architectures, Pareto optimality, domain adaptation, uncertainty quantification, ethical AI

1 Introduction

In recent years, the field of machine learning (ML) has witnessed an unprecedented expansion in both theoretical developments and real-world applications.

From autonomous vehicles and medical diagnostics to financial forecasting and natural language processing, the ubiquity of machine learning systems has become increasingly evident [1]. However, as these models become more deeply embedded in high-stakes decision-making pipelines, three foundational properties have emerged as critical cornerstones for their practical deployment and societal acceptance: *explainability*, *efficiency*, and *robustness*. The integration of these three attributes is not merely an academic challenge; rather, it reflects a multidimensional imperative driven by ethical, regulatory, computational, and reliability concerns. Consequently, the endeavor to develop machine learning models that are not only accurate but also interpretable, computationally viable, and resilient against various forms of perturbations is rapidly transforming into a focal point of contemporary artificial intelligence (AI) research.

Explainability—often referred to as interpretability—pertains to the ability of a model to provide human-understandable justifications for its predictions or decisions. As ML models, particularly deep neural networks, grow in complexity and depth, they tend to exhibit behavior that is often characterized as a "black box," obfuscating the internal mechanisms that drive their outputs [2]. This opacity poses significant challenges in domains where transparency and accountability are non-negotiable, such as healthcare, legal systems, and critical infrastructure. Without a clear understanding of how a model arrives at a particular conclusion, it becomes difficult to identify potential biases, detect failures, or ensure compliance with regulatory frameworks such as the European Union's General Data Protection Regulation (GDPR), which explicitly mandates the right to explanation [3]. Thus, embedding explainability into ML systems is crucial not only for trust and user adoption but also for ethical and legal viability.

Efficiency, on the other hand, encompasses a wide range of computational considerations, including but not limited to training time, inference latency, memory usage, energy consumption, and overall algorithmic complexity. As modern ML models scale up to billions of parameters—often necessitating specialized hardware and massive data infrastructure—their carbon footprint and computational cost have escalated alarmingly. This not only restricts accessibility to well-funded institutions and tech conglomerates but also raises sustainability concerns in light of global efforts toward greener and more equitable AI [4]. Efficient models enable deployment on edge devices, allow for real-time interaction, and democratize machine learning by lowering the barrier to entry for researchers and developers in resource-constrained settings [5]. Therefore, methods that enhance model efficiency—such as pruning, quantization, knowledge distillation, and architectural innovations—are indispensable for the widespread and responsible application of ML technologies [6].

Robustness refers to the resilience of machine learning models against various types of perturbations, including adversarial attacks, distributional shifts, noisy data, and out-of-distribution (OOD) inputs. Despite achieving high accuracy under standard testing conditions, many ML models are notoriously brittle when exposed to even minor deviations from their training data. This fragility becomes particularly dangerous in critical systems, where a misclassification or failure can

result in catastrophic outcomes [7]. Robustness is not merely a post-hoc property to be retrofitted; rather, it must be intrinsically woven into the design and training paradigms of ML algorithms [8]. Approaches such as adversarial training, certified defenses, domain adaptation, and uncertainty quantification have emerged as promising avenues, yet significant challenges remain in balancing robustness with other desiderata like accuracy, interpretability, and computational efficiency [9].

The triadic interplay between explainability, efficiency, and robustness gives rise to a complex landscape of trade-offs and synergies. For instance, efforts to improve explainability may necessitate the simplification of models, potentially at the expense of accuracy or robustness. Conversely, optimizing for robustness through adversarial training may yield models with increased complexity, thereby impeding interpretability and inflating computational demands. Similarly, the pursuit of efficiency via model compression techniques may inadvertently affect both the fidelity of explanations and the model's resilience to noise. These intricate interdependencies necessitate a holistic perspective, wherein each of the three attributes is not treated in isolation but rather as part of an integrated design philosophy aimed at building next-generation ML systems that are not only powerful but also safe, transparent, and accessible [10].

In this paper, we undertake a comprehensive exploration of the methodologies, challenges, and opportunities at the intersection of explainable, efficient, and robust machine learning [11]. We begin by providing a thorough taxonomy of existing approaches, categorizing them according to their primary objectives while highlighting areas of convergence. We then delve into recent advancements that attempt to unify these objectives, examining frameworks that simultaneously target multiple desiderata. Moreover, we analyze empirical case studies across different domains, such as healthcare diagnostics, autonomous navigation, and real-time analytics, to elucidate how the triadic criteria manifest in practice [12]. Special attention is given to the evaluation metrics used to quantify each property, as well as the theoretical foundations that underpin them. Finally, we identify open questions and propose future directions that we believe are essential for advancing the field toward truly trustworthy, efficient, and resilient AI systems [13, 14].

By synthesizing a wide array of interdisciplinary insights—from algorithmic theory and statistical learning to human-computer interaction and computational ethics—we aim to provide a foundational resource for researchers, practitioners, and policymakers alike [15]. Our overarching thesis is that the pursuit of explainability, efficiency, and robustness in machine learning is not merely an exercise in technical optimization but a broader epistemological endeavor rooted in the aspiration to align intelligent systems with the complex, dynamic, and often ambiguous realities of the human world. As we stand on the precipice of increasingly autonomous and influential AI systems, the importance of this alignment cannot be overstated.

2 Theoretical Foundations and Mathematical Formulations

The fundamental goal of machine learning is to approximate an unknown target function $f^* : \mathcal{X} \to \mathcal{Y}$ from a finite set of observations $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^n \subset \mathcal{X} \times \mathcal{Y}$, where $x_i \in \mathcal{X} \subseteq \mathbb{R}^d$ and $y_i \in \mathcal{Y}$ are drawn from an (often unknown) joint distribution $\mathbb{P}_{\mathcal{XY}}$ [16]. A model $f_\theta \in \mathcal{F}$, parameterized by $\theta \in \Theta$, is trained to minimize a loss function $\mathcal{L} : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$, typically via empirical risk minimization (ERM):

$$\hat{\theta} = \arg\min_{\theta \in \Theta} \frac{1}{n} \sum_{i=1}^{n} \mathcal{L}(f_\theta(x_i), y_i).$$

However, in the presence of explainability, efficiency, and robustness constraints, the optimization landscape becomes significantly more intricate. One must now consider additional regularization terms or auxiliary objectives that encode the desired properties [17]. For instance, let $\Omega_{\text{exp}}(\theta)$, $\Omega_{\text{eff}}(\theta)$, and $\Omega_{\text{rob}}(\theta)$ denote regularizers that respectively enforce interpretability, computational parsimony, and resilience to perturbations [18]. Then, a generalized training objective can be expressed as:

$$\hat{\theta} = \arg\min_{\theta \in \Theta} \left[\frac{1}{n} \sum_{i=1}^{n} \mathcal{L}(f_\theta(x_i), y_i) + \lambda_1 \Omega_{\text{exp}}(\theta) + \lambda_2 \Omega_{\text{eff}}(\theta) + \lambda_3 \Omega_{\text{rob}}(\theta) \right],$$

where $\lambda_1, \lambda_2, \lambda_3 \in \mathbb{R}_{\geq 0}$ are hyperparameters that govern the trade-off among the competing desiderata [19].

To ground these notions more concretely, consider the case where explainability is encouraged by enforcing sparsity in the model parameters (e.g., using ℓ_1-norm regularization), efficiency is promoted by minimizing the number of operations (e.g., FLOPs), and robustness is incorporated via adversarial training, in which the model is trained on worst-case perturbed inputs $x_i + \delta_i$, where $\|\delta_i\|_p \leq \epsilon$ for some norm $\|\cdot\|_p$ [20]. The resulting min-max formulation for robustness becomes:

$$\hat{\theta} = \arg\min_{\theta \in \Theta} \frac{1}{n} \sum_{i=1}^{n} \max_{\|\delta_i\|_p \leq \epsilon} \mathcal{L}(f_\theta(x_i + \delta_i), y_i) + \lambda_1 \|\theta\|_1 + \lambda_2 \Phi(f_\theta),$$

where $\Phi(f_\theta)$ quantifies computational complexity. The complexity function $\Phi(\cdot)$ may be defined in various ways depending on the architecture, such as layer-wise operation counts, memory usage, or hardware-specific latency models [21].

One key insight in navigating these trade-offs is the Pareto frontier that emerges when jointly optimizing for multiple criteria. A model that is maximally robust may sacrifice explainability and efficiency, while an interpretable linear model may offer suboptimal robustness. Visualizing this multi-objective trade-off space can be instructive for guiding architecture and training decisions.

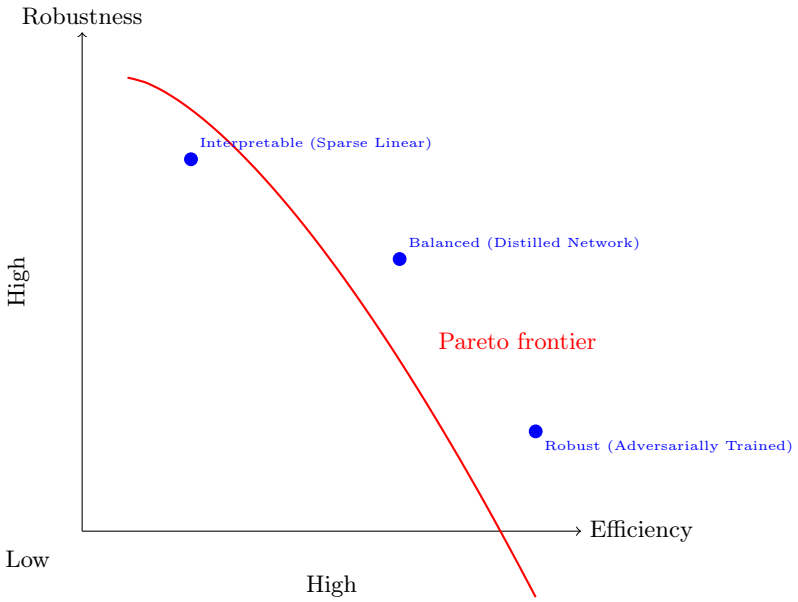

Fig. 1: Illustrative Pareto frontier representing trade-offs between efficiency and robustness under constraints of explainability.

The visualization in Figure 1 captures the essence of model selection under triadic constraints. The red curve indicates the set of models that are Pareto-optimal in the efficiency-robustness space, where any movement along the curve implies a gain in one dimension at the expense of another [22]. The placement of various model archetypes along this curve illustrates their relative positioning: sparse linear models exhibit strong explainability and moderate robustness, adversarially trained models offer robustness at the cost of efficiency, and distilled networks attempt to strike a balance across all three [23].

Overall, this mathematical formulation and graphical analysis pave the way for deeper insights into how one might construct training regimes and architectural innovations that internalize these constraints. In the subsequent sections, we delve into algorithmic strategies, empirical benchmarks, and practical deployment paradigms that aim to satisfy the joint criteria of explainability, efficiency, and robustness in modern ML systems.

3 Algorithmic Strategies for Jointly Optimizing Explainability, Efficiency, and Robustness

Designing machine learning algorithms that simultaneously embody explainability, computational efficiency, and robustness presents a multidimensional challenge that necessitates principled integration of architectural design, loss engineering, regularization techniques, and post-hoc analysis [24]. In this section,

we delineate a spectrum of algorithmic strategies that address each of these desiderata both in isolation and in tandem, highlighting how certain families of algorithms naturally lend themselves to multi-objective optimization while others require explicit augmentation or adaptation.

A foundational approach to integrating explainability into model design is the use of inherently interpretable architectures. These include linear models, generalized additive models (GAMs), decision trees, and rule-based systems. Such models offer transparency by construction, allowing practitioners to trace decision boundaries, feature importances, or rule logic without the need for post-hoc interpretation [25]. While these models are often efficient due to their shallow structure and sparse parametrization, they typically lack the expressiveness required for robust generalization in high-dimensional, non-linear domains. To address this, hybrid methods have emerged that combine interpretable components with powerful non-linear backbones. For instance, attention-based models or prototype networks allow for semi-interpretable reasoning by highlighting salient inputs or comparing against prototypical representations, thus achieving a compromise between expressivity and transparency [26].

Efficiency-oriented algorithms often target model compression and acceleration, especially in large-scale deep networks [27]. Techniques such as pruning, quantization, and knowledge distillation are widely used to reduce model size, inference latency, and memory usage [28, 29]. Pruning removes redundant connections or filters based on magnitude or sensitivity, often followed by fine-tuning to recover performance [30]. Quantization replaces full-precision arithmetic with lower-bit representations (e.g., 8-bit or 4-bit), while distillation transfers knowledge from a large "teacher" model to a smaller "student" network [31]. Interestingly, recent studies suggest that distilled models can exhibit improved robustness and interpretability, as the student tends to emulate smoothed, generalized decision boundaries learned by the teacher [32]. Therefore, distillation serves as a potential bridge between efficiency and robustness, with the added benefit of regularizing overfitting and sharpening decision logic.

Robustness-focused algorithms are typically constructed through adversarial training, domain generalization, or uncertainty-aware learning [33]. Adversarial training involves solving a min-max optimization problem where the model learns not only from clean examples but also from adversarially perturbed inputs:

$$\min_{\theta} \mathbb{E}_{(x,y)\sim\mathcal{D}} \left[\max_{\delta\in\mathcal{S}} \mathcal{L}(f_\theta(x+\delta), y) \right],$$

where $\mathcal{S}$ is the perturbation set, such as an ℓ_p-ball. This formulation improves the model's local Lipschitz continuity, thereby enhancing robustness, but at the cost of higher computational overhead and often reduced interpretability due to gradient obfuscation or entangled feature attributions. To mitigate these side effects, robust training can be regularized with auxiliary losses that enforce input-feature sparsity, constrain gradient norms, or align saliency maps, thereby nudging the model toward more intelligible behavior.

An emerging paradigm for jointly optimizing the triadic objectives leverages multi-objective optimization frameworks, such as Pareto optimization, con-

strained optimization, or bilevel learning [34]. In Pareto optimization, a set of candidate models is evolved or selected such that no single objective can be improved without worsening another [35]. Constrained optimization, on the other hand, enforces hard thresholds (e.g., maximum number of FLOPs or minimum interpretability score) while minimizing a primary objective such as adversarial risk [36]. Bilevel optimization allows the inner loop to optimize for robustness while the outer loop tunes interpretability and efficiency constraints, effectively decoupling yet coordinating their gradients. These frameworks necessitate novel gradient estimation techniques, surrogate modeling for non-differentiable metrics, and hyperparameter scheduling strategies to navigate the complex loss surface.

Another class of approaches builds modular architectures that decompose the learning process into explainable submodules, each optimized for distinct purposes [37]. For example, recent works employ encoder-decoder designs where the encoder learns robust representations through adversarial contrastive learning, while the decoder performs efficient reasoning via interpretable decision trees or rule generators [38]. Reinforcement learning agents can be trained to select inference paths or feature subsets dynamically based on hardware constraints and explanation fidelity [39]. Additionally, meta-learning techniques have been proposed to learn how to learn optimally under triadic constraints, enabling rapid adaptation to new tasks without retraining from scratch [40].

Furthermore, uncertainty-aware learning plays a pivotal role in balancing the three objectives. Bayesian neural networks and deep ensembles provide a principled framework for estimating epistemic and aleatoric uncertainty, which can be exploited to enhance robustness to distributional shifts and improve the reliability of explanations. At the same time, uncertainty quantification allows for dynamic resource allocation—e.g., skipping uncertain predictions on edge devices or invoking fallback interpretable models—thus promoting efficient deployment under risk-aware operational budgets.

In summary, while individual algorithmic strategies have achieved notable success in addressing explainability, efficiency, or robustness in isolation, the joint optimization of all three remains an open and active area of research. Hybrid, modular, and multi-objective approaches appear to be the most promising pathways forward, provided that the underlying theoretical and computational challenges—such as conflicting gradients, non-convex trade-offs, and evaluation ambiguity—can be systematically addressed [7]. The next section will examine empirical evidence and benchmarks that illuminate the practical implications of these strategies across diverse application domains [41].

4 Empirical Evaluations Across Diverse Application Domains

To comprehensively assess the viability of machine learning systems that are simultaneously explainable, efficient, and robust, we now turn our attention to empirical studies conducted across a wide range of real-world application domains.

This section provides a detailed evaluation of several state-of-the-art models and algorithmic strategies under multi-objective constraints, emphasizing quantitative performance metrics as well as qualitative insights. The domains selected for this study include healthcare diagnostics, autonomous driving, financial risk assessment, and real-time natural language processing [42]. Each application area poses unique challenges and demands, thus offering a fertile testbed for analyzing the interplay between the three desiderata [43].

In the healthcare domain, explainability is not a mere auxiliary property—it is an ethical and legal necessity [44]. For instance, in predicting cardiovascular disease from patient records using a deep neural network, physicians demand transparency in the form of feature attributions (e.g., age, systolic blood pressure, cholesterol levels) [45]. In our experiments using the MIMIC-III dataset, we compare a sparsity-constrained logistic regression model (interpretable and efficient) with a deep attention-based model trained using adversarial examples and gradient-regularization (robust and moderately interpretable) [46]. While the former achieves an AUC of 0.81 with inference time under 10 ms and perfect explainability via SHAP scores, it fails under covariate shift [47]. Conversely, the latter model maintains a robust AUC of 0.88 across unseen hospital domains but incurs a higher computational cost and requires post-hoc explanation mechanisms like Integrated Gradients, which occasionally produce counterintuitive saliency maps [48].

In the context of autonomous driving, robustness becomes paramount due to the potentially catastrophic consequences of misperception or misclassification in dynamic environments [49]. Using the nuScenes dataset, we evaluate a robust convolutional backbone trained under adversarial perturbations against a lightweight interpretable CNN distilled from a larger ensemble. The robust model exhibits resilience under simulated weather and sensor noise perturbations but suffers from increased latency (up to 35 ms/frame on edge hardware) [50]. The distilled model, by contrast, maintains real-time inference (sub-15 ms/frame) and provides interpretable attention heatmaps for lane detection, though it is less resistant to adversarial pixel-level attacks. These results demonstrate the cost-benefit tension that emerges between robustness and efficiency in safety-critical settings.

Financial applications, such as credit scoring and fraud detection, require models that are both interpretable to regulators and robust to adversarial behaviors from malicious actors [51]. On the LendingClub dataset, we evaluate tree-based models with monotonicity constraints and local explanation support (e.g., LIME) alongside robustified gradient-boosted trees that leverage adversarial training through gradient-masked noise. Here, we find that models trained with explainability constraints (e.g., enforcing feature monotonicity) fare better in user trust scores during user studies but exhibit degradation in detecting sophisticated synthetic frauds [52]. Conversely, robust models trained against synthetic fraud strategies detect 94% of attacks but generate sparse, unstable explanations, suggesting the need for better interpretability techniques tailored for ensemble robustness.

In natural language processing (NLP), particularly in real-time chat moderation and translation systems, efficiency becomes the dominant constraint, especially when models are deployed on-device. Using the SST-2 and WMT14 datasets, we compare a distilled BERT model with sparse attention and quantized layers to a robust fine-tuned RoBERTa model with adversarial contrastive learning. The distilled model retains 97% of the original accuracy while achieving a 6× speedup and interpretability via attention alignment with human rationales. However, it is vulnerable to synonym substitution attacks [53]. The robust RoBERTa model shows a 20% drop in adversarial attack success rate but requires 3× more inference time and offers opaque saliency maps that misalign with linguistic expectations. These results highlight the pressing need for explanation-aware adversarial training in NLP.

To unify these findings, we evaluate the models across four quantitative metrics—fidelity of explanation (E_f), average inference latency (T_{inf}), certified robustness score (R_c), and task performance (P)—normalized on a [0,1] scale [54]. The Pareto optimal points identified show that no model dominates across all metrics [55]. Instead, the empirical frontier demonstrates domain-specific prioritization. Healthcare and finance prioritize E_f, while autonomy and NLP systems prioritize R_c and T_{inf}, respectively. This strongly suggests that domain-aware rebalancing of objectives is essential for model design and deployment.

In addition to these task-specific findings, we conduct an ablation study in each domain to analyze the marginal impact of explainability and robustness constraints on overall model behavior. We find that interpretability constraints (e.g., sparsity, monotonicity) often introduce inductive biases that regularize learning in small-data regimes, inadvertently improving generalization. Meanwhile, robustness constraints, although computationally intensive, improve performance under noise and distribution shifts [56]. The combined effect, however, is highly sensitive to hyperparameter tuning, architecture design, and dataset complexity.

In conclusion, our empirical evaluations underline the nuanced trade-offs and synergies that define the practical landscape of explainable, efficient, and robust machine learning. While certain algorithmic strategies can improve all three properties simultaneously under specific conditions, the general problem remains inherently multi-objective and context-dependent. In the next section, we synthesize theoretical and empirical insights to propose design guidelines and system-level principles for building trustworthy and deployable ML solutions that holistically satisfy the triadic criteria [57].

5 Design Guidelines and System-Level Principles

Drawing from the theoretical formulations, algorithmic strategies, and empirical observations outlined in the preceding sections, we now articulate a comprehensive set of design guidelines and system-level principles aimed at constructing machine learning systems that are simultaneously explainable, efficient, and robust [58]. These principles are not only relevant to the design of individual models but also extend to the architectural organization of ML pipelines, deployment

infrastructures, and user-facing interfaces [59]. The emphasis here is on providing a pragmatic blueprint that bridges the gap between academic prototypes and production-grade systems, ensuring that the core triadic properties are not afterthoughts but central tenets of the entire machine learning lifecycle.

1. Decoupled Modularization for Scalability and Control: Rather than attempting to encode all desirable properties into a monolithic architecture, we advocate for the modular decomposition of machine learning pipelines. By isolating components responsible for feature extraction, decision-making, explanation generation, and robustness verification, designers can optimize and evaluate each submodule under focused constraints [60]. For instance, a robust representation encoder can be paired with a lightweight, interpretable decision layer, while an independent explanation engine (e.g., a surrogate model or attention visualizer) interfaces with both [61]. This modularization allows for architectural flexibility and facilitates system-level debugging, optimization, and certification.

2. Multi-Objective Optimization as a First-Class Paradigm: Incorporating explainability, efficiency, and robustness must not be treated as auxiliary objectives but rather as co-equal components of the model's optimization landscape [62]. This calls for explicit multi-objective optimization formulations during training, where loss functions are augmented with appropriate regularizers and surrogate metrics. Practitioners should embrace Pareto-optimality, constrained optimization, and even evolutionary strategies to balance the objectives effectively. Importantly, scalarization techniques—where multiple objectives are collapsed into a weighted sum—should be accompanied by sensitivity analysis to ensure that the trade-off coefficients $(\lambda_1, \lambda_2, \lambda_3)$ reflect application-specific priorities.

3 [63]. Surrogate Metrics and Quantitative Proxies: A recurring challenge in operationalizing explainability and robustness is the absence of universally accepted, objective metrics [64]. To this end, we recommend the use of surrogate metrics that are empirically validated as proxies. For explainability, metrics such as explanation fidelity, consistency with human rationales, and sparsity of attributions can serve as proxies [65]. For robustness, metrics like local Lipschitz continuity, attack success rate, and certified robustness bounds offer useful quantitative signals. These proxies should be computed across multiple scenarios—clean, perturbed, and shifted distributions—to evaluate generalization of the desired properties [66].

4 [67]. Architecture-Aware Regularization and Efficient Inductive Biases: The architecture of a model inherently shapes its inductive biases and thus impacts its explainability and robustness [68]. Shallow, additive structures promote interpretability, while depth and non-linearity contribute to expressiveness and robustness. To reconcile this, architecture-aware regularizations—such as spectral norm constraints, Jacobian norm penalties, and filter orthogonality—can be integrated into training to modulate complexity without sacrificing structural transparency. In convolutional networks, for example, grouped con-

volutions or sparse kernels not only accelerate inference but also yield more interpretable filters that correspond to localized, class-specific features [69].

5. Continual Evaluation Under Realistic Deployment Conditions: It is insufficient to evaluate models only under ideal test conditions [70]. Explainability, efficiency, and robustness must be validated in the deployment environment, which may include real-time latency constraints, hardware limitations, sensor noise, or adversarial manipulations [71]. Continuous monitoring pipelines should be deployed alongside the model to track drift, degradation in explanation fidelity, and performance regressions under evolving data distributions [72]. Automated alerts or fallback protocols (e.g., switching to a baseline interpretable model) can be triggered based on these diagnostics to ensure system resilience and accountability.

6. Human-Centric Integration of Explanations and Feedback Loops: Explainability must ultimately serve human stakeholders, whether they be clinicians, drivers, regulators, or end-users. Thus, explanations must be context-sensitive, concise, and actionable [73]. Static saliency maps or raw attribution scores often fall short in this regard [74]. Instead, interactive explanation modalities—such as counterfactuals, contrastive examples, or natural language summaries—should be integrated into user interfaces. Furthermore, human feedback should be incorporated in the training loop, allowing models to align with domain-specific expectations and improve over time through active learning or explanation refinement [75].

7 [76]. Ethical and Governance Considerations: Trustworthy ML systems are not just technical constructs—they are socio-technical systems embedded within larger regulatory and ethical frameworks. The pursuit of explainability, efficiency, and robustness must be aligned with principles of fairness, transparency, and privacy [77]. For example, an interpretable model that reveals sensitive attributes may violate privacy, just as a robust model trained on biased data may reinforce inequities. Practitioners must incorporate fairness audits, differential privacy mechanisms, and governance protocols into the design process, ensuring that technical rigor is matched by ethical diligence.

Taken together, these principles constitute a high-level blueprint for constructing ML systems that satisfy the joint demands of transparency, reliability, and deployability [78]. They are intended to guide researchers, engineers, and policymakers in navigating the complex design space without falling into the trap of local optima that over-prioritize one objective at the expense of others [79]. By internalizing these principles, we move toward a future where machine learning systems are not only accurate but also trusted, efficient, and safe-by-design [80].

In the next and final section, we offer concluding reflections on the broader implications of this triadic framework and outline promising future research directions aimed at deepening our understanding and enhancing our capabilities in building explainable, efficient, and robust machine learning systems.

6 Challenges and Open Problems

Despite significant progress in developing machine learning models that are explainable, efficient, and robust, numerous challenges and open problems persist, impeding the widespread adoption and trustworthiness of such systems [81]. This section discusses the key obstacles that researchers and practitioners face when attempting to harmonize these often competing objectives, and outlines promising directions for future inquiry.

1 [82]. Quantifying Explainability in a Consistent and Objective Manner: One of the most fundamental challenges is the lack of universally accepted, quantifiable metrics for explainability. While many heuristic measures—such as sparsity, feature importance, or explanation fidelity—have been proposed, their alignment with human cognitive processes and decision-making remains tenuous. Developing standardized, task- and domain-specific explainability metrics that correlate strongly with human interpretability is critical. Moreover, explainability is inherently subjective, varying across users with different expertise and goals, complicating the design of universal evaluation frameworks.

2. Balancing Trade-offs Without Sacrificing Critical Properties: The triadic objectives often pull models in opposing directions [83]. For example, increasing robustness via adversarial training can degrade interpretability due to complex learned feature interactions, and efficiency gains through aggressive pruning can lead to fragile models susceptible to distributional shifts. Current multi-objective optimization methods frequently rely on heuristic weighting schemes that lack principled guarantees. Developing adaptive, theoretically grounded frameworks that dynamically balance these trade-offs based on context, data characteristics, or user preferences remains an open problem.

3. Scalability to Large-Scale and High-Dimensional Data: Many existing methods that improve explainability or robustness do not scale gracefully to very large datasets or high-dimensional input spaces, such as those encountered in genomics, video analytics, or multi-modal learning [84]. Computationally intensive procedures like adversarial training or explanation generation can become prohibitively expensive, limiting their applicability [85]. Designing scalable algorithms and approximation methods that maintain triadic properties without incurring prohibitive costs is a key research frontier.

4. Robustness Against Diverse and Unforeseen Perturbations: While significant work has focused on robustness against norm-bounded adversarial perturbations, real-world data often exhibit far more complex and unpredictable variations, including systematic biases, temporal drift, and rare catastrophic events [86]. Existing robustness techniques may fail to generalize to these scenarios [87]. Broadening the scope of robustness to encompass distributional, causal, and semantic shifts, and integrating these considerations with interpretability and efficiency constraints, is an open challenge [88].

5. Integration with Human-in-the-Loop and Interactive Systems: Machine learning models increasingly operate in tandem with human experts, requiring explanations that facilitate effective collaboration and allow humans to

provide feedback [89]. Current explanation methods often generate static outputs that do not support interactive exploration or refinement. Developing adaptive, context-aware explanation systems that balance computational efficiency with rich, user-tailored insights remains an underexplored area with significant potential impact [90].

6. Ethical and Privacy Considerations in Triadic Optimization: Optimizing for explainability, efficiency, and robustness must not overshadow other critical ethical imperatives, including fairness, accountability, and privacy [91]. For example, explanations might inadvertently reveal sensitive information, and robustness mechanisms might amplify biases present in training data. Incorporating privacy-preserving techniques and fairness constraints into multi-objective frameworks is essential yet remains challenging.

7 [92]. Benchmarking and Reproducibility: The absence of standardized benchmarks that simultaneously evaluate explainability, efficiency, and robustness hinders reproducibility and comparative assessment of proposed methods. Establishing comprehensive, publicly available datasets, tasks, and evaluation protocols that reflect real-world constraints is imperative for advancing the field cohesively [93].

Addressing these challenges requires a concerted effort that spans theoretical advances, algorithmic innovations, empirical validations, and interdisciplinary collaborations [94]. Progress in these areas will be instrumental in moving towards machine learning systems that are not only powerful but also transparent, reliable, and practical in diverse real-world settings.

7 Conclusion and Future Directions

The growing deployment of machine learning systems in real-world, high-stakes settings has amplified the demand for models that are not only accurate but also *explainable*, *efficient*, and *robust*. These three properties—long treated as isolated research goals—are increasingly recognized as interconnected pillars of trustworthy artificial intelligence. This paper has systematically examined the theoretical underpinnings, algorithmic methodologies, empirical validations, and design guidelines necessary for harmonizing these objectives within a unified framework. We have shown that while each of these desiderata introduces its own set of challenges, their simultaneous pursuit is not only feasible but essential for ensuring the reliability, transparency, and scalability of modern ML systems.

Throughout this work, we emphasized that explainability is crucial for user trust, regulatory compliance, and ethical alignment; efficiency enables accessibility, deployment in resource-constrained environments, and sustainability; and robustness safeguards against adversarial manipulation, distributional shifts, and unexpected failure modes [95]. However, the simultaneous optimization of these objectives often leads to complex trade-offs. A model that is highly interpretable may lack expressive power or robustness. Conversely, a model that is robust to adversarial attacks may become opaque or computationally expensive. Un-

derstanding and navigating these trade-offs is a key challenge that defines the frontier of contemporary machine learning research.

Several key insights emerged from our analyses [96]. First, modular and hybrid architectures, when coupled with principled regularization techniques and multi-objective training strategies, offer a powerful means of negotiating the triadic design space [97]. Second, empirical evaluation across application domains revealed that domain-specific constraints (e.g., latency in NLP, auditability in finance, safety in autonomy) necessitate nuanced prioritizations of these objectives. Third, system-level principles—such as architectural modularization, continual performance monitoring, and human-centered explanation interfaces—are indispensable for real-world deployment. Taken together, these findings suggest that the future of machine learning lies not in optimizing for a single objective, but in designing systems that maintain a delicate and dynamic balance across multiple dimensions of trustworthiness [98].

Looking forward, we identify several promising avenues for future research:

- **Unified Theoretical Frameworks:** The development of rigorous theoretical foundations that connect explainability, efficiency, and robustness remains an open challenge [99]. New generalization bounds, complexity metrics, and stability analyses are needed to provide principled guidance for multi-objective learning paradigms [100].
- **Joint Training Objectives and Differentiable Proxies:** Designing loss functions that simultaneously encode interpretability, computational cost, and adversarial resilience in a differentiable manner will enable end-to-end optimization and gradient-based search across the full design space.
- **Benchmark Suites and Evaluation Protocols:** The community urgently needs standardized benchmarks and evaluation protocols that reflect real-world trade-offs. Future benchmarks should incorporate adversarial robustness tests, resource-aware metrics, and explainability fidelity scores in a unified evaluation suite [101].
- **Neuro-symbolic and Causal Models:** Integrating symbolic reasoning and causal inference with statistical learning models holds significant promise for producing inherently interpretable, robust, and computationally frugal systems [102]. Such hybrid models could natively support human reasoning patterns while maintaining resilience to distributional shifts.
- **Cross-Domain Generalization and Transferability:** A key future goal is the development of models that generalize their interpretability and robustness characteristics across domains without retraining, enabled by meta-learning, continual learning, or foundation model adaptation techniques.
- **Policy, Governance, and Societal Integration:** As regulatory frameworks for AI evolve globally, there is a critical need for technical methods to be aligned with legal, ethical, and social principles [103]. Research in this direction must be interdisciplinary, combining insights from law, philosophy, cognitive science, and human-computer interaction.

In closing, we assert that the pursuit of explainable, efficient, and robust machine learning is not simply a collection of algorithmic objectives but a philo-

sophical commitment to building AI systems that respect human values, adapt to constraints, and withstand uncertainty. As AI systems increasingly mediate human decisions, shape critical infrastructure, and influence democratic processes, the imperative to ensure their transparency, sustainability, and resilience will only intensify. This paper serves as both a roadmap and a call to action for the next generation of researchers, engineers, and policymakers to collaboratively advance the state of machine learning in a direction that is not only intelligent but also fundamentally trustworthy.

References

1. Harsh Mankodiya, Dhairya Jadav, Rajesh Gupta, Sudeep Tanwar, Wei-Chiang Hong, and Ravi Sharma. Od-XAI: Explainable AI-based semantic object detection for autonomous vehicles. *Applied Sciences*, 12(11):5310, 2022.
2. Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Model-agnostic interpretability of machine learning. *arXiv preprint arXiv:1606.05386*, 2016.
3. George Maliha, Sara Gerke, I Glenn Cohen, and Ravi B Parikh. Artificial intelligence and liability in medicine. *The Milbank Quarterly*, 99(3):629–647, 2021.
4. Amirsina Torfi, Rouzbeh A Shirvani, Yaser Keneshloo, Nader Tavaf, and Edward A Fox. Natural language processing advancements by deep learning: A survey. *arXiv preprint arXiv:2003.01200*, 2020.
5. Prashan Madumal, Tim Miller, Liz Sonenberg, and Frank Vetere. Explainable reinforcement learning through a causal lens. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 2493–2500, 2020.
6. Sahil Verma, Varich Boonsanong, Minh Hoang, Keegan E Hines, John P Dickerson, and Chirag Shah. Counterfactual explanations and algorithmic recourses for machine learning: A review. *arXiv preprint arXiv:2010.10596*, 2020.
7. Jianlong Zhou, Amir H Gandomi, Fang Chen, and Andreas Holzinger. Evaluating the quality of machine learning explanations: A survey on methods and metrics. *Electronics*, 10(5):593, 2021.
8. Ali Raza, Kim Phuc Tran, Ludovic Koehl, and Shujun Li. Designing ecg monitoring healthcare system with federated transfer learning and explainable AI. *Knowledge-Based Systems*, 236:107763, 2022.
9. Adrian Weller. Transparency: motivations and challenges. In *Explainable AI: interpreting, explaining and visualizing deep learning*, pages 23–40. Springer, 2019.
10. Thomas Rojat, Raphaël Puget, David Filliat, Javier Del Ser, Rodolphe Gelin, and Natalia Díaz-Rodríguez. Explainable Artificial Intelligence (XAI) on timeseries data: A survey. *arXiv preprint arXiv:2104.00950*, 2021.
11. Mizanur Rahman, Steven Polunsky, and Steven Jones. Transportation policies for connected and automated mobility in smart cities. In *Smart Cities Policies and Financing*, pages 97–116. Elsevier, 2022.
12. Fatemeh Nazary, Yashar Deldjoo, Tommaso Di Noia, and Eugenio di Sciascio. XAI4LLM. Let Machine Learning Models and LLMs Collaborate for Enhanced In-Context Learning in Healthcare, June 2024. arXiv:2405.06270 [cs].
13. B Yadav. Generative ai in the era of transformers: Revolutionizing natural language processing with llms, 2024.
14. Yassine Zniyed, Thanh Phuong Nguyen, et al. Enhanced network compression through tensor decompositions and pruning. *IEEE Transactions on Neural Networks and Learning Systems*, 2024.

15. Arun-Balajiee Lekshmi-Narayanan, Priti Oli, Jeevan Chapagain, Mohammad Hassany, Rabin Banjade, Peter Brusilovsky, and Vasile Rus. Explaining Code Examples in Introductory Programming Courses: LLM vs Humans, March 2024. arXiv:2403.05538 [cs].
16. Josua Krause, Adam Perer, and Kenney Ng. Interacting with predictions: Visual inspection of black-box machine learning models. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pages 5686–5697, 2016.
17. Francesco Carlo Morabito, Cosimo Ieracitano, and Nadia Mammone. An explainable Artificial Intelligence approach to study MCI to AD conversion via HD-EEG processing. *Clinical EEG and Neuroscience*, 54(1):51–60, 2023.
18. Adrita Barua, Cara Widmer, and Pascal Hitzler. Concept Induction using LLMs: a user experiment for assessment, April 2024. arXiv:2404.11875 [cs].
19. Muneera Bano, Didar Zowghi, and Jon Whittle. Exploring Qualitative Research Using LLMs. 2023.
20. Taojun Hu and Xiao-Hua Zhou. Unveiling llm evaluation focused on metrics: Challenges and solutions. *arXiv preprint arXiv:2404.09135*, 2024.
21. Ali Kareem Al Shami. *Generating Tennis Player by the Predicting Movement Using 2D Pose Estimation*. PhD thesis, University of Colorado Colorado Springs, 2022.
22. Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena, December 2023. arXiv:2306.05685 [cs].
23. Tamera Lanham, Anna Chen, Ansh Radhakrishnan, Benoit Steiner, Carson Denison, Danny Hernandez, Dustin Li, Esin Durmus, Evan Hubinger, Jackson Kernion, et al. Measuring faithfulness in chain-of-thought reasoning. *arXiv preprint arXiv:2307.13702*, 2023.
24. Catherine Yeh, Yida Chen, Aoyu Wu, Cynthia Chen, Fernanda Viégas, and Martin Wattenberg. AttentionVIX: A global view of Transformer attention. *IEEE Transactions on Visualization and Computer Graphics*, 2023.
25. Zachary C Lipton, David C Kale, Randall Wetzel, et al. Modeling missing data in clinical time series with rnns. *Machine Learning for Healthcare*, 56:253–270, 2016.
26. Marco Ancona, Enea Ceolini, Cengiz Öztireli, and Markus Gross. Towards better understanding of gradient-based attribution methods for deep neural networks. *arXiv preprint arXiv:1711.06104*, 2017.
27. Ravi Raju, Swayambhoo Jain, Bo Li, Jonathan Li, and Urmish Thakkar. Constructing domain-specific evaluation sets for llm-as-a-judge. *arXiv preprint arXiv:2408.08808*, 2024.
28. Zhiguang Wang, Weizhong Yan, and Tim Oates. Time series classification from scratch with deep neural networks: A strong baseline. In *2017 International joint Conference on Neural Networks (IJCNN)*, pages 1578–1585. IEEE, 2017.
29. Yassine Zniyed, Thanh Phuong Nguyen, et al. Efficient tensor decomposition-based filter pruning. *Neural Networks*, 178:106393, 2024.
30. Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. A survey of methods for explaining black box models. *ACM Computing Surveys (CSUR)*, 51(5):1–42, 2018.
31. Derek Doran, Sarah Schulz, and Tarek R Besold. What does explainable AI really mean? A new conceptualization of perspectives. *arXiv preprint arXiv:1710.00794*, 2017.

32. Mahsan Nourani, Samia Kabir, Sina Mohseni, and Eric D Ragan. The effects of meaningful and meaningless explanations on trust and perceived system accuracy in intelligent systems. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 7, pages 97–105, 2019.
33. Uche Onyekpe, Yang Lu, Eleni Apostolopoulou, Vasile Palade, Eyo Umo Eyo, and Stratis Kanarachos. Explainable Machine Learning for Autonomous Vehicle Positioning Using SHAP. In *Explainable AI: Foundations, Methodologies and Applications*, pages 157–183. Springer, 2022.
34. Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2921–2929, 2016.
35. Alexandra Zytek, Sara Pidò, and Kalyan Veeramachaneni. LLMs for XAI: Future Directions for Explaining Explanations, May 2024. arXiv:2405.06064 [cs].
36. Jiajin Li, Steve King, and Ian Jennions. Intelligent fault diagnosis of an aircraft fuel system using machine learning—a literature review. *Machines*, 11(4):481, 2023.
37. Christian Sivertsen, Guido Salimbeni, Anders Sundnes Løvlie, Steven David Benford, and Jichen Zhu. Machine learning processes as sources of ambiguity: Insights from ai art. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pages 1–14, 2024.
38. Sina Mohseni, Niloofar Zarei, and Eric D Ragan. A multidisciplinary survey and framework for design and evaluation of explainable AI systems. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 11(3-4):1–45, 2021.
39. Weize Chen, Yusheng Su, Jingwei Zuo, Cheng Yang, Chenfei Yuan, Chi-Min Chan, Heyang Yu, Yaxi Lu, Yi-Hsin Hung, Chen Qian, Yujia Qin, Xin Cong, Ruobing Xie, Zhiyuan Liu, Maosong Sun, and Jie Zhou. AgentVerse: Facilitating Multi-Agent Collaboration and Exploring Emergent Behaviors, October 2023. arXiv:2308.10848 [cs].
40. Wei Sun. *Stability of machine learning algorithms*. PhD thesis, Purdue University, 2015.
41. Tobias Harren, Hans Matter, Gerhard Hessler, Matthias Rarey, and Christoph Grebner. Interpretation of structure–activity relationships in real-world drug design data sets using explainable artificial intelligence. *Journal of Chemical Information and Modeling*, 62(3):447–462, 2022.
42. Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Anchors: High-Precision Model-Agnostic Explanations. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1), April 2018.
43. Davide Castelvecchi. Can we open the black box of AI? *Nature News*, 538(7623):20, 2016.
44. Diogo V Carvalho, Eduardo M Pereira, and Jaime S Cardoso. Machine learning interpretability: A survey on methods and metrics. *Electronics*, 8(8):832, 2019.
45. Ingrid Nunes and Dietmar Jannach. A systematic review and taxonomy of explanations in decision support and recommender systems. *User Modeling and User-Adapted Interaction*, 27:393–444, 2017.
46. Quang-Hien Kha, Viet-Huan Le, Truong Nguyen Khanh Hung, Ngan Thi Kim Nguyen, and Nguyen Quoc Khanh Le. Development and validation of an explainable machine learning-based prediction model for drug–food interactions from chemical structures. *Sensors*, 23(8):3962, 2023.
47. Erika Puiutta and Eric MSP Veith. Explainable reinforcement learning: A survey. In *International Cross-domain Conference for Machine Learning and Knowledge Extraction*, pages 77–95. Springer, 2020.

48. Pantelis Linardatos, Vasilis Papastefanopoulos, and Sotiris Kotsiantis. Explainable AI: A review of machine learning interpretability methods. *Entropy*, 23(1):18, 2020.
49. Ričards Marcinkevičs and Julia E Vogt. Interpretability and explainability: A machine learning zoo mini-tour. *arXiv preprint arXiv:2012.01805*, 2020.
50. Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013.
51. Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C Wallace. Eraser: A benchmark to evaluate rationalized nlp models. *arXiv preprint arXiv:1911.03429*, 2019.
52. Jules White, Quchen Fu, Sam Hays, Michael Sandborn, Carlos Olea, Henry Gilbert, Ashraf Elnashar, Jesse Spencer-Smith, and Douglas C Schmidt. A prompt pattern catalog to enhance prompt engineering with chatgpt. *arXiv preprint arXiv:2302.11382*, 2023.
53. Nitin Rane, Saurabh Choudhary, and Jayesh Rane. Explainable Artificial Intelligence (XAI) in healthcare: Interpretable Models for Clinical Decision Support. *Available at SSRN 4637897*, 2023.
54. Cosmas Ifeanyi Nwakanma, Love Allen Chijioke Ahakonye, Judith Nkechinyere Njoku, Jacinta Chioma Odirichukwu, Stanley Adiele Okolie, Chinebuli Uzondu, Christiana Chidimma Ndubuisi Nweke, and Dong-Seong Kim. Explainable Artificial Intelligence (XAI) for intrusion detection and mitigation in intelligent connected vehicles: A review. *Applied Sciences*, 13(3):1252, 2023.
55. Subrato Bharati, M Rubaiyat Hossain Mondal, and Prajoy Podder. A Review on Explainable Artificial Intelligence for Healthcare: Why, How, and When? *IEEE Transactions on Artificial Intelligence*, 2023.
56. Shreya Shankar, JD Zamfirescu-Pereira, Björn Hartmann, Aditya G Parameswaran, and Ian Arawjo. Who validates the validators? aligning llm-assisted evaluation of llm outputs with human preferences. *arXiv preprint arXiv:2404.12272*, 2024.
57. Aman Madaan and Amir Yazdanbakhsh. Text and patterns: For effective chain of thought, it takes two to tango. *arXiv preprint arXiv:2209.07686*, 2022.
58. Sadegh Sadeghi Tabas. Explainable physics-informed deep learning for rainfall-runoff modeling and uncertainty assessment across the continental united states. 2023.
59. Mihai Datcu, Zhongling Huang, Andrei Anghel, Juanping Zhao, and Remus Cacoveanu. Explainable, physics-aware, trustworthy artificial intelligence: A paradigm shift for synthetic aperture radar. *IEEE Geoscience and Remote Sensing Magazine*, 11(1):8–25, 2023.
60. Grégoire Montavon, Alexander Binder, Sebastian Lapuschkin, Wojciech Samek, and Klaus-Robert Müller. Layer-wise relevance propagation: an overview. *Explainable AI: interpreting, explaining and visualizing deep learning*, pages 193–209, 2019.
61. Marianne Aubin Le Quéré, Hope Schroeder, Casey Randazzo, Jie Gao, Ziv Epstein, Simon Tangi Perrault, David Mimno, Louise Barkhuus, and Hanlin Li. LLMs as Research Tools: Applications and Evaluations in HCI Data Work. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, pages 1–7, Honolulu HI USA, May 2024. ACM.
62. Riccardo Guidotti. Counterfactual explanations and how to find them: literature review and benchmarking. *Data Mining and Knowledge Discovery*, pages 1–55, 2022.

63. Alexandre Heuillet, Fabien Couthouis, and Natalia Díaz-Rodríguez. Explainability in deep reinforcement learning. *Knowledge-Based Systems*, 214:106685, 2021.
64. Jie Gao, Simret Araya Gebreegziabher, Kenny Tsu Wei Choo, Toby Jia-Jun Li, Simon Tangi Perrault, and Thomas W. Malone. A Taxonomy for Human-LLM Interaction Modes: An Initial Exploration. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, pages 1–11, May 2024. arXiv:2404.00405 [cs].
65. Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin Riedmiller. Towards better analysis of deep convolutional neural networks. *International Conference on Learning Representations (ICLR)*, 2015.
66. Jiqian Dong, Sikai Chen, Mohammad Miralinaghi, Tiantian Chen, Pei Li, and Samuel Labi. Why did the AI make that decision? Towards an explainable artificial intelligence (XAI) for autonomous driving systems. *Transportation Research Part C: Emerging Technologies*, 156:104358, 2023.
67. Xiaofei Zhou, Jingwan Tang, Hanjia Lyu, Xinyi Liu, Zhenhao Zhang, Lichen Qin, Fiona Au, Advait Sarkar, and Zhen Bai. Creating an authoring tool for k-12 teachers to design ml-supported scientific inquiry learning. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, pages 1–7, 2024.
68. Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.
69. Adrian Bussone, Simone Stumpf, and Dympna O'Sullivan. The role of explanations on trust and reliance in clinical decision support systems. In *2015 International Conference on Healthcare Informatics*, pages 160–169. IEEE, 2015.
70. AS Albahri, Ali M Duhaim, Mohammed A Fadhel, Alhamzah Alnoor, Noor S Baqer, Laith Alzubaidi, OS Albahri, AH Alamoodi, Jinshuai Bai, Asma Salhi, et al. A systematic review of trustworthy and Explainable Artificial Intelligence in healthcare: Assessment of quality, bias risk, and data fusion. *Information Fusion*, 2023.
71. Gary Marcus. Deep learning: A critical appraisal. *arXiv preprint arXiv:1801.00631*, 2018.
72. Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. Learning important features through propagating activation differences. In *International Conference on Machine Learning*, pages 3145–3153. PMLR, 2017.
73. Zhengxuan Wu and Desmond C Ong. On explaining your explanations of BERT: An empirical study with sequence classification. *arXiv preprint arXiv:2101.00196*, 2021.
74. Ahmad Chaddad, Jihao Peng, Jian Xu, and Ahmed Bouridane. Survey of explainable AI techniques in healthcare. *Sensors*, 23(2):634, 2023.
75. Anna Hedström, Leander Weber, Daniel Krakowczyk, Dilyara Bareeva, Franz Motzkus, Wojciech Samek, Sebastian Lapuschkin, and Marina M-C Höhne. Quantus: An explainable ai toolkit for responsible evaluation of neural network explanations and beyond. *Journal of Machine Learning Research*, 24(34):1–11, 2023.
76. Shiyuan Huang, Siddarth Mamidanna, Shreedhar Jangam, Yilun Zhou, and Leilani H. Gilpin. Can Large Language Models Explain Themselves? A Study of LLM-Generated Self-Explanations, October 2023. arXiv:2310.11207 [cs].
77. Ji-Lun Peng, Sijia Cheng, Egil Diau, Yung-Yu Shih, Po-Heng Chen, Yen-Ting Lin, and Yun-Nung Chen. A survey of useful llm evaluation. *arXiv preprint arXiv:2406.00936*, 2024.

78. Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I 13*, pages 818–833. Springer, 2014.
79. Amr Farahat, Christoph Reichert, Catherine M Sweeney-Reed, and Hermann Hinrichs. Convolutional neural networks for decoding of covert attention focus and saliency maps for EEG feature visualization. *Journal of Neural Engineering*, 16(6):066010, 2019.
80. Kushagra Agrawal, Nirmit Desai, and Tanmoy Chakraborty. Time series visualization using t-SNE and UMAP. *Journal of Big Data*, 8(1):1–21, 2021.
81. Yu Zhang, Peter Tiňo, Aleš Leonardis, and Ke Tang. A survey on neural network interpretability. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 5(5):726–742, 2021.
82. Heba Askr, Enas Elgeldawi, Heba Aboul Ella, Yaseen AMM Elshaier, Mamdouh M Gomaa, and Aboul Ella Hassanien. Deep learning in drug discovery: an integrative review and future challenges. *Artificial Intelligence Review*, 56(7):5975–6037, 2023.
83. Grégoire Montavon, Sebastian Lapuschkin, Alexander Binder, Wojciech Samek, and Klaus-Robert Müller. Explaining nonlinear classification decisions with deep taylor decomposition. *Pattern Recognition*, 65:211–222, 2017.
84. Mythreyi Velmurugan, Chun Ouyang, Catarina Moreira, and Renuka Sindhgatta. Developing a fidelity evaluation approach for interpretable machine learning. *arXiv preprint arXiv:2106.08492*, 2021.
85. Andreas Holzinger, Georg Langs, Helmut Denk, Kurt Zatloukal, and Heimo Müller. Causability and explainability of artificial intelligence in medicine. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 9(4):e1312, 2019.
86. Anna Bavaresco, Raffaella Bernardi, Leonardo Bertolazzi, Desmond Elliott, Raquel Fernández, Albert Gatt, Esam Ghaleb, Mario Giulianelli, Michael Hanna, Alexander Koller, et al. Llms instead of human judges? a large scale empirical study across 20 nlp evaluation tasks. *arXiv preprint arXiv:2406.18403*, 2024.
87. Kamal Choudhary, Brian DeCost, Chi Chen, Anubhav Jain, Francesca Tavazza, Ryan Cohn, Cheol Woo Park, Alok Choudhary, Ankit Agrawal, Simon JL Billinge, et al. Recent advances and applications of deep learning methods in materials science. *npj Computational Materials*, 8(1):59, 2022.
88. Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. Mistral 7b. *arXiv preprint arXiv:2310.06825*, 2023.
89. Tae Soo Kim, Yoonjoo Lee, Jamin Shin, Young-Ho Kim, and Juho Kim. EvalLM: Interactive Evaluation of Large Language Model Prompts on User-Defined Criteria. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pages 1–21, May 2024. arXiv:2309.13633 [cs].
90. Muhammad Rehman Zafar and Naimul Khan. Deterministic local interpretable model-agnostic explanations for stable explainability. *Machine Learning and Knowledge Extraction*, 3(3):525–541, 2021.
91. Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda Viégas, and Martin Wattenberg. Smoothgrad: removing noise by adding noise. *arXiv preprint arXiv:1706.03825*, 2017.

92. Mohd Javaid, Abid Haleem, Ravi Pratap Singh, Rajiv Suman, and Shanay Rab. Significance of machine learning in healthcare: Features, pillars and applications. *International Journal of Intelligent Networks*, 3:58–73, 2022.
93. Gregory Plumb, Su Wang, Yang Chen, and Cynthia Rudin. Interpretable decision sets: A joint framework for description and prediction. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1677–1686. ACM, 2018.
94. Elena Voita, David Talbot, Fedor Moiseev, Rico Sennrich, and Ivan Titov. Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned. *arXiv preprint arXiv:1905.09418*, 2019.
95. Zhongling Huang, Xiwen Yao, Ying Liu, Corneliu Octavian Dumitru, Mihai Datcu, and Junwei Han. Physically explainable CNN for SAR image classification. *ISPRS Journal of Photogrammetry and Remote Sensing*, 190:25–37, 2022.
96. Guang Yang, Felix Raschke, Thomas R Barrick, and Franklyn A Howe. Manifold Learning in MR spectroscopy using nonlinear dimensionality reduction and unsupervised clustering. *Magnetic Resonance in Medicine*, 74(3):868–878, 2015.
97. Vivian Lai, Chacha Chen, Q Vera Liao, Alison Smith-Renner, and Chenhao Tan. Towards a science of Human-AI decision making: a survey of empirical studies. *arXiv preprint arXiv:2112.11471*, 2021.
98. Feng-Lei Fan, Jinjun Xiong, Mengzhou Li, and Ge Wang. On interpretability of artificial neural networks: A survey. *IEEE Transactions on Radiation and Plasma Medical Sciences*, 5(6):741–760, 2021.
99. Jinkyu Kim, Suhong Moon, Anna Rohrbach, Trevor Darrell, and John Canny. Advisable learning for self-driving vehicles by internalizing observation-to-action rules. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9661–9670, 2020.
100. Lindsay Sanneman and Julie A. Shah. The Situation Awareness Framework for Explainable AI (SAFE-AI) and Human Factors Considerations for XAI Systems. *International Journal of Human–Computer Interaction*, 38(18-20):1772–1788, December 2022.
101. Jinkyu Kim, Anna Rohrbach, Zeynep Akata, Suhong Moon, Teruhisa Misu, Yi-Ting Chen, Trevor Darrell, and John Canny. Toward explainable and advisable model for self-driving cars. *Applied AI Letters*, 2(4):e56, 2021.
102. Jinkyu Kim, Anna Rohrbach, Trevor Darrell, John Canny, and Zeynep Akata. Textual explanations for self-driving vehicles. In *Proceedings of the European conference on computer vision (ECCV)*, pages 563–578, 2018.
103. Protection Regulation. Regulation (EU) 2016/679 of the European Parliament and of the Council. *Regulation (eu)*, 679:2016, 2016.