Decision Making and Analysis
https://ojs.luminescience.cn/DMA

LUMINESCENCE
https://luminescience.cn/

Review

Decision-oriented explainable artificial intelligence: a PDR-based review of methods, applications, and emerging frontiers

Jiangshan Zhu

School of Management, Hangzhou Dianzi University, Hangzhou, 310018, China

Corresponding to: Jiangshan Zhu, Email: Jiangshan.Zhu@hdu.edu.cn

Abstract: Explainable artificial intelligence (XAI) is increasingly used in consequential decisions, but method selection still requires evidence about predictive reliability, explanatory fidelity, and stakeholder usefulness. This latent Dirichlet allocation (LDA)-assisted structured review used the OpenAlex Works metadata application programming interface (API) to retrieve 4,200 English-language records published from 2018 to 13 July 2026. Sequential DOI, exact-title, and fuzzy-title deduplication retained 3,166 records; title, abstract, and keyword screening retained 694; eligibility assessment retained 666; and 666 records entered the final document-term matrix. Candidate LDA models with $k = 5 - 16$ were compared using c_v and u_mass coherence, conventional perplexity, matched-topic stability, and Jensen–Shannon separation. The selected $k = 8$ solution achieved c_v = 0.5226, stability = 0.6809, and mean topic separation = 0.5824. Manual inspection of the top terms and documents identified themes concerning robust counterfactual generation, responsible and regulated decision support, interpretable clinical risk prediction, industrial, cybersecurity, and infrastructure XAI, causal and actionable algorithmic recourse, visual and multimodal medical explanation, clinical XAI adoption, trust, and workflow, and human-centered XAI evaluation and taxonomy. These evidence-derived themes are integrated with the adopted Predictive-Descriptive-Relevance (PDR) framework as a decision-oriented synthesis lens. The review contributes a reproducible literature map, a critical method comparison, a stakeholder-centered cross-domain matrix, and a research roadmap for causal, robust, uncertainty-aware, multimodal, generative, foundation-model XAI.

Keywords: PDR framework, LDA topic modeling, Intrinsic interpretability, Post-hoc explainability, Decision support, XAI

1. Introduction

Deep learning has expanded the predictive reach of artificial intelligence, but its use in consequential decisions has intensified concerns about accountability, fairness, safety, and justified trust [1-2]. Visual and feature-level explanations can make model behavior inspectable, yet an explanation that appears plausible is not necessarily faithful, stable, or useful to the person who must act on it [3-4]. The practical question is therefore not simply whether an explanation can be produced, but whether its evidence is sufficiently reliable and relevant for a specified decision process.

This review distinguishes interpretability from explainability. Interpretability denotes model or representation properties that are understandable by

Received: Jun. 18, 2026; Revised: Jul. 13, 2026; Accepted: Jul. 24, 2026; Published: Aug. 14, 2026

DOI: https://doi.org/10.55976/dma.42026159087-99

design; explainability denotes post-hoc evidence or communication intended to clarify an already trained model. Accordingly, methods are organized as intrinsically interpretable models and post-hoc explainability methods. The Predictive-Descriptive-Relevance (PDR) framework is adopted from Murdoch et al. and reinterpreted as a decision-oriented synthesis lens; it is not presented as a newly invented theory [5-6].

Influential surveys have clarified explainable artificial intelligence (XAI) terminology, taxonomies, program history, and recurring technical challenges [2, 7-9]. Recent reviews have also mapped explainability for large language models and generative artificial intelligence (AI) [10-11]. However, these contributions do not jointly combine an empirical topic map, a PDR-based method comparison, a stakeholder-oriented application matrix, and a deployment research roadmap. A detailed comparison of these organizing choices is provided in Table S1 in the Supplementary Material.

The comparison indicates that the present contribution lies in integration rather than in proposing a new XAI taxonomy or a new PDR theory. The empirical topic model makes the literature map auditable; PDR then connects model reliability, explanation fidelity, and stakeholder usefulness; and the application matrix tests whether those criteria change across decision settings. This combination supports method selection and reporting decisions that algorithm-centered surveys alone do not resolve.

The review makes four contributions:

(1) a reproducible OpenAlex retrieval, staged deduplication, screening, preprocessing, and LDA protocol;

(2) a topic-informed PDR synthesis that treats PDR as an adopted decision lens;

(3) a critical comparison of intrinsic and post-hoc methods across fidelity, stability, actionability, scalability, and decision context; and

(4) a cross-domain stakeholder matrix and structured roadmap spanning causal, uncertainty-aware, robust, concept-based, multimodal, generative, and foundation-model XAI.

2. Literature retrieval, latent Dirichlet allocation (LDA) topic modeling, and review methodology

2.1 Review design

The study is a LDA-assisted structured review rather than a full Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) systematic review. LDA is used to identify recurrent lexical themes in the eligible metadata corpus, while critical narrative synthesis is used to compare methods, applications, and decision implications. The design follows the probabilistic topic-modeling framework introduced by Blei et al. and Griffiths and Steyvers [12-13]. The algorithm does not replace substantive review; it provides an auditable evidence map for the subsequent PDR synthesis.

2.2 Data source and search strategy

Metadata were retrieved from the OpenAlex Works application programming interface (API) on 13 July 2026 for English-language records published from 1 January 2018 through 13 July 2026. Earlier seminal papers were retained separately for conceptual grounding and were not modeled. OpenAlex was the only scholarly metadata source queried; Web of Science, Scopus, IEEE Xplore, ACM Digital Library, PubMed, and other databases are not reported as searched.

Twelve relevance-ranked OpenAlex queries were organized into A AND B, A AND C, and A AND B AND C families, where A represented XAI anchors, B represented methods/frontiers, and C represented decisions or applications. Each query was capped at 350 records; OpenAlex meta-counts and downloaded counts were preserved. The method anchors included SHapley Additive exPlanations (SHAP) and Local Interpretable Model-agnostic Explanations (LIME). The exact query strings were:

(1) explainable artificial intelligence SHAP LIME counterfactual explanation; (2) interpretable machine learning feature attribution rule sets local surrogate models; (3) explainable artificial intelligence knowledge graph causal explanation uncertainty robustness; (4) explainable artificial intelligence large language models foundation models multimodal generative AI; (5) explainable artificial intelligence decision support healthcare finance law education; (6) interpretable machine learning high stakes decision making clinical decision support; (7) explainable AI environmental science cybersecurity agriculture risk management; (8) human centered explainable artificial intelligence stakeholder decision governance; (9) explainable artificial intelligence SHAP decision support healthcare; (10) counterfactual explanation algorithmic recourse decision making; (11) interpretable machine learning causal uncertainty decision support; (12) explainable artificial intelligence foundation model decision support.

2.3 Screening and corpus construction

The API returned 4,200 records. Deduplication was performed in a fixed sequence: normalized DOI (3,222 retained), exact normalized title (3,169), and high-similarity normalized title at a SequenceMatcher threshold of 0.965 (3,166). Title/abstract/keyword screening required a substantive XAI anchor plus a method/frontier or decision/application focus and retained 694 records. Eligibility required an adequate English abstract, an allowed scholarly record type, and substantive XAI focus; 666 records remained. Explicit exclusions covered

particle and high-energy physics, generic prediction or optimization without substantive explanation, editorials, notices, corrupted metadata, and incidental mentions of explainability.

After preprocessing and removal of empty vectors, 666 records entered LDA; seven seminal papers were retained outside the model. Figure 1 reports these counts and distinguishes retrieval, deduplication, screening, eligibility, and modeling.

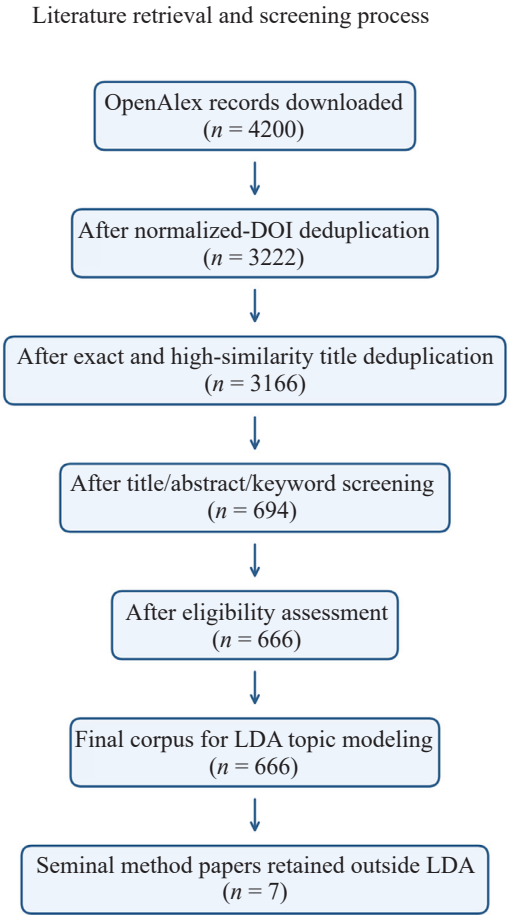

Figure 1. Literature retrieval and screening process for the LDA-assisted structured review

2.4 Text preprocessing and document-term matrix

The text modeling combined titles, abstracts, and OpenAlex keywords; broad OpenAlex concept labels were retained only as metadata and were not modeled. Processing included lowercase normalization, removal of URLs and non-letter characters, whitespace tokenization, WordNet noun-then-verb lemmatization, standard and project-specific stopwords, and deterministic phrase mapping for terms such as large_language_model, counterfactual_explanation, algorithmic_recourse, feature_attribution, concept_bottleneck, integrated_gradients, and human_centered. CountVectorizer used min_df = 3, max_df = 0.55, and keep_n/max_features = 2,200. The final document-term matrix contained 666 rows and 2,200 terms.

2.5 Topic-number selection and LDA estimation

Candidate models with $k = 5 - 16$ were estimated using batch variational Bayes for 15 iterations with seed 42. Default symmetric document-topic and topic-word priors (1/k) were used. Additional seeds 123 and 2026 were used to assess matched-topic stability. Diagnostics included c_v and u_mass coherence, conventional perplexity (lower is better), per-token log likelihood, matched-topic cosine stability, and mean pairwise Jensen–Shannon distance. Coherence metrics followed established evaluation studies [14-17]. The selected $k = 8$ solution achieved c_v = 0.5226, u_mass = -1.9445, perplexity = 1136.03, stability = 0.6809, and mean topic separation = 0.5824. Although smaller models had lower perplexity and higher stability, $k = 8$ had the highest weighted selection score and clearer substantive separation.

Figure 2 shows the full coherence and model-fit comparison prior to the selected solution.

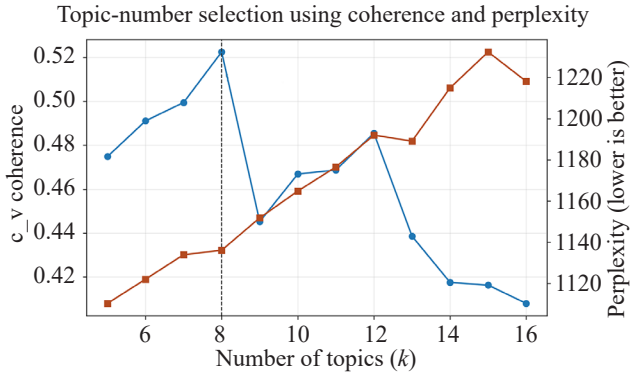

Figure 2. Topic-number selection using coherence and model-fit diagnostics

2.6 Topic visualization, interpretation, and manual coding

Intertopic geometry was calculated from Jensen–Shannon distances between normalized topic-word distributions and projected using metric multidimensional scaling; this avoids treating raw topic weights as Euclidean principal component analysis (PCA) coordinates. The visualization follows the interpretive intent of LDAvis [18].

Topic interpretation used a single, author-assisted computational coding pass. For each selected topic, the top 30 terms and top 20 documents by posterior probability were inspected; off-topic records were excluded before

the final rerun; labels were assigned only after semantic inspection; and two to four representative papers were selected from the verified documents. No independent double screening, inter-rater agreement, or Cohen's kappa is claimed.

Figure 3 presents the final eight-topic distance map; the manually interpreted themes are reported in Table S2 in the Supplementary Material.

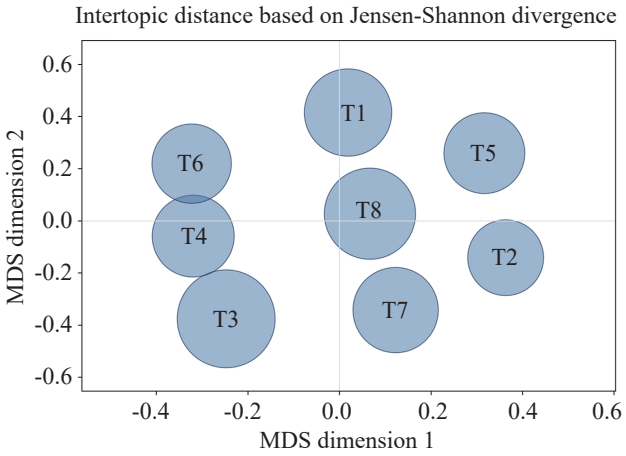

Figure 3. Intertopic-distance visualization for the selected eight-topic LDA model

2.7 Integration with the PDR framework

The LDA topics are treated as an empirical map rather than a substitute for conceptual synthesis. Method-centered topics inform descriptive accuracy; clinical, governance, security, and domain-use topics inform relevance and predictive risk; and counterfactual themes connect explanation to feasible action. Sections 3–7 therefore use the PDR framework to compare methods, identify stakeholder-dependent explanation needs, and translate topic gaps into a research roadmap.

3. PDR framework

3.1 Motivation and definition

PDR evaluates an XAI system through three distinct questions: whether the predictor is reliable, whether the explanatory artifact faithfully describes the predictor, and whether that artifact is relevant to a defined audience and decision. Keeping these dimensions separate prevents a persuasive visualization from being mistaken for evidence of model quality and prevents predictive accuracy from being mistaken for decision legitimacy [4-5, 19].

3.2 Predictive accuracy

Predictive accuracy concerns generalization to the target population, not only performance on a convenient test split. Evaluation should include calibration, subgroup errors, distribution shift, and robustness to reasonable perturbations. Explanations built on a poorly calibrated or unstable predictor cannot serve as strong decision evidence, even when the explanatory procedure is internally consistent.

3.3 Descriptive accuracy

Descriptive accuracy concerns fidelity to the model behavior that an explanation purports to describe. Relevant evidence includes local surrogate fit, perturbation fidelity, sensitivity, randomization checks, and consistency across functionally similar models [20-21]. The relationship between accuracy and interpretability is context dependent rather than universal. In structured, high-stakes tabular tasks, interpretable models can match or exceed black-box performance and should be tested directly; in image, speech, and large-language tasks, expressive models may retain predictive advantages [6, 22].

3.4 Relevance

Relevance asks whether an explanation improves performance on a real task for an intended stakeholder. A clinician may need calibrated risk estimates and concise case evidence; an applicant may need contestable reasons and feasible avenues for recourse; an auditor may need global constraints, stability tests, and traceable documentation. Human satisfaction alone is insufficient if the explanation increases automation bias or does not improve decision quality.

3.5 Mapping methodological categories to PDR

Intrinsically interpretable models expose decision structure through coefficients, smooth effects, monotonicity, trees, or rules. Post-hoc explainability methods analyze a trained predictor through attribution, local surrogates, examples, concepts, counterfactuals, semantic paths, or visual evidence. Intrinsic structure can improve auditability but still requires validation; post-hoc methods can be practical for expressive models but risk rationalizing them rather than faithfully explaining them. Figure 4 shows how task context, PDR evidence, method selection, stakeholder use, and monitoring form a single evaluation loop.

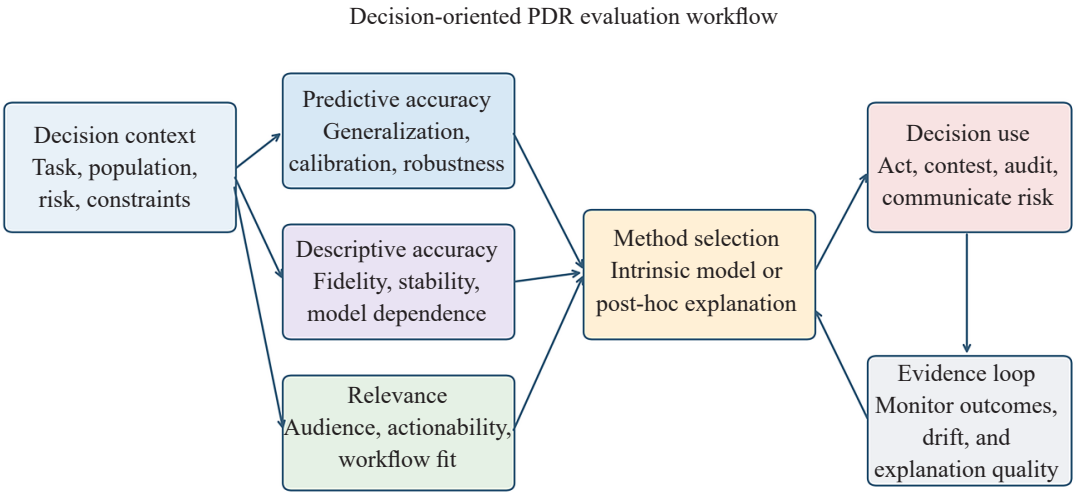

Figure 4. Decision-oriented PDR framework for XAI method evaluation

3.6 Operationalization, assessment, and reporting

A PDR report should combine predictive evidence (test, subgroup, calibration, and shift performance), descriptive evidence (fidelity, stability, sensitivity, and model dependence), and relevance evidence (task outcome, comprehension, time, contestability, and downstream harm). The intended user, decision, available actions, and institutional constraints should be specified before selecting the explanation. This ordering turns XAI evaluation from a visualization exercise into a decision-support protocol.

4. Interpretable methods

Figure 5 organizes representative approaches into intrinsically interpretable models and post-hoc explainability methods. The categories indicate where interpretability enters the system; they do not guarantee that an output is faithful or useful. Each method must still be evaluated against the PDR requirements of the deployment context.

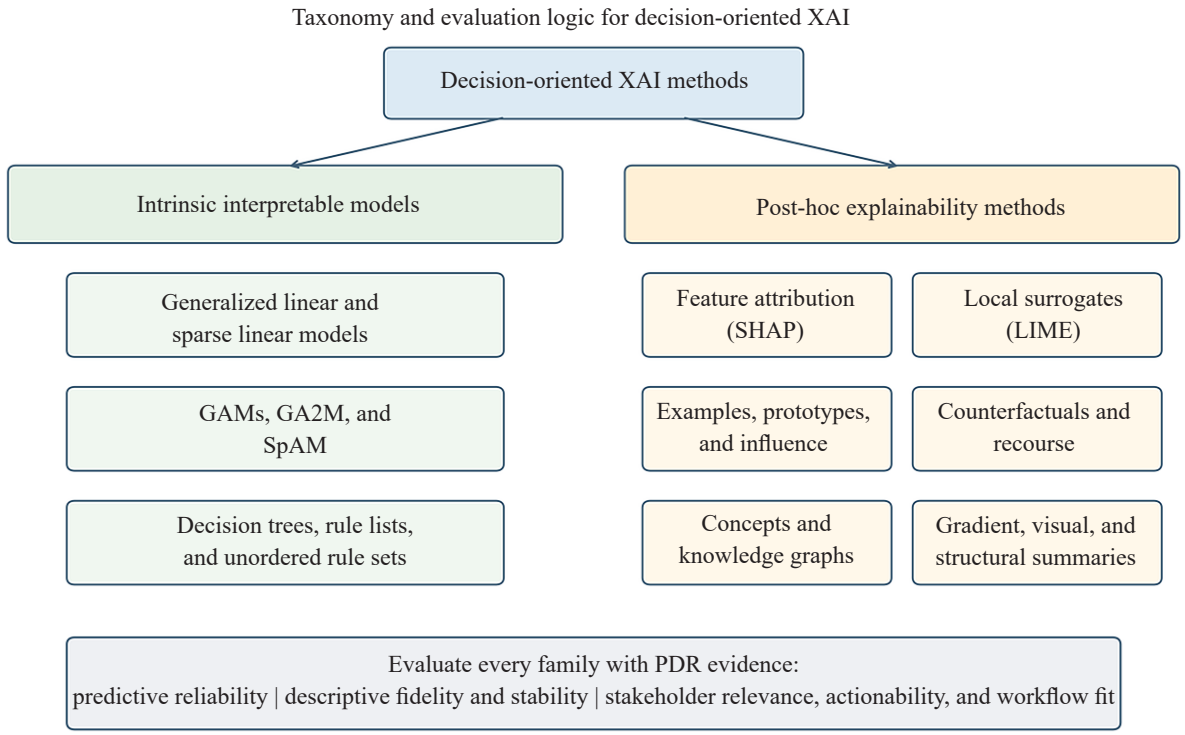

Figure 5. Decision-oriented taxonomy of XAI methods under the PDR framework

4.1 Intrinsic interpretable models

4.1.1 Linear and logistic regression

A linear model represents the response as an additive function of observed features:

$$y = \beta_0 + \sum_{j=1}^{p} \beta_j x_j + \varepsilon.$$

Here y is the response, x_j is the j-th feature, β_0 is the intercept, and β_j is the corresponding coefficient. The term ε denotes the unexplained error, and p is the number of features. Interpreting the coefficients assumes linearity on the response scale, appropriate feature specification, and adequate control of dependence and confounding factors. A large coefficient does not automatically indicate causality or decision relevance.

For a binary outcome, logistic regression applies the logit link:

$$\log\left(\frac{P(y=1|\ x)}{1-P(y=1|\ x)}\right) = \beta_0 + \sum_{j=1}^{p} \beta_j x_j.$$

Here $P(y=1\mid x)$ is the conditional probability, and each β_j represents the change in the log-odds for a one-unit increase in x_j, holding other features fixed. Logistic coefficients are globally interpretable, but nonlinear relationships, interactions, collinearity, and distribution shift can make a seemingly simple model misleading in practice.

4.1.2 Additive and sparse additive models

A generalized additive model (GAM) replaces fixed slopes with smooth univariate functions:

$$g(E[y|\ x]) = \beta_0 + \sum_{j=1}^{p} f_j\left(x_j\right).$$

The link function g connects the conditional mean to the additive predictor, and the partial-effect curve f_j is associated with feature j. Interpretability depends on additivity, smoothness control, and support within the observed feature range. Pairwise extensions add selected interactions:

$$g(E[y|\ x]) = \beta_0 + \sum_{j=1}^{p} f_j\left(x_j\right) + \sum_{j<k} f_{jk}\left(x_j, x_k\right).$$

The f_{jk} terms encode selected two-way interactions. Generalized additive models with pairwise interactions (GA2M/GAMI) generally constrain the number and complexity of interactions so that the resulting surfaces remain auditable. Sparse additive models (SpAM) impose sparsity across component functions [23-24]. These models often provide a strong PDR compromise for tabular decisions because global effects can be audited without assuming that every relationship is linear.

4.1.3 Decision trees, rule lists, and rule sets

Classification and regression trees (CART) control tree complexity by penalizing empirical risk with tree size:

$$\min_T \{R(T) + \alpha|\ T|\}.$$

T is a candidate tree, $R(T)$ is its empirical loss, $|\ T\ |$ is the number of terminal leaves, and α penalizes complexity. Small trees are simulable and support audit trails, but axis-aligned splits can be unstable and are often inefficient for approximating smooth relationships [25-26].

Ordered rule lists apply rules sequentially, so each decision depends on earlier else conditions [27]. Unordered rule sets, such as Repeated Incremental Pruning to Produce Error Reduction (RIPPER), CN2, and Interpretable Decision Sets, allow independent rules to fire and can represent multiple sufficient pathways [28-30]. Their main risks are overlap, conflicting coverage, and a rapid loss of simplicity as the rule count grows.

4.2 Post-hoc explainability methods

4.2.1 Prototypes, criticisms, and influence functions

Maximum mean discrepancy (MMD)-Critic compares a dataset X with a prototype set P in a reproducing-kernel Hilbert space (RKHS):

$$MMD^2(P, X) = \left\| \frac{1}{|\ X\ |} \sum_{x \in X} \phi(x) - \frac{1}{|\ P\ |} \sum_{p \in P} \phi(p) \right\|_H^2.$$

X is the observed sample, P is the selected prototype subset, ϕ maps observations into the RKHS, and the norm measures the discrepancy between the mean embeddings of X and P. Prototypes summarize well-covered regions, whereas criticisms reveal poorly represented ones. The results depend on the choice of kernel and do not imply causal importance [31].

Influence functions approximate the change in test loss at a test point z_{test} when a training point z is upweighted:

$$I_{\text{up, loss}}\left(z, z_{\text{test}}\right) = -\nabla_\theta L\left(z_{\text{test}}, \hat{\theta}\right)^\top H_{\hat{\theta}}^{-1} \nabla_\theta L(z, \hat{\theta}).$$

Here, L is the loss, $\hat{\theta}$ is the fitted parameter vector, and $H_{\hat{\theta}}$ is the Hessian of the empirical risk at the solution. This approximation assumes differentiability and a well-conditioned Hessian. However, non-convexity, damping choices, and inverse-Hessian approximation can affect rankings [32].

4.2.2 SHAP and feature attribution

For a feature set N with M features, the SHAP value for feature i is:

$$\phi_i = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(M-|S|-1)!}{M!}[v(S \cup \{i\}) - v(S)],$$

where, N is the complete feature set, S is a coalition excluding i, M is the number of features, and $v(S)$ is the value assigned to coalition S under a specified conditional or interventional semantics. While SHAP offers axiomatic consistency, factors such as dependence-sensitive semantics, background data, approximation, and feature representation can materially change the result [33]. Therefore, stable attributions require explicit conditioning assumptions and sensitivity analysis.

4.2.3 LIME and local surrogate models

LIME selects a simple surrogate model g around instance x by optimizing:

$$\xi(x) = \text{argmin}_{g \in G}\{L(f, g, \pi_x) + \Omega(g)\},$$

where f is the black-box predictor, G is the class of interpretable models, and π_x weights samples by their proximity to x. The term L measures local disagreement, while $\Omega(g)$ penalizes complexity. Although LIME is model-agnostic and easy to interpret, factors such as neighborhood definition, sampling distribution, representation, kernel width, and random seed can lead to varying explanations [34]. Thus, model agnosticism comes at the cost of sensitivity to neighborhood design.

4.2.4 Counterfactual explanations and algorithmic recourse

A common counterfactual objective seeks a feasible alternative x_{cf}:

$$x_{cf} = \text{argmin}_{x' \in A}\left\{d\left(x, x'\right) + \lambda L\left(f\left(x'\right), y'\right)\right\},$$

where d penalizes departure from the original x, L measures the loss relative to the desired outcome y', and λ balances proximity with target attainment. The set A represents feasibility and actionability constraints. A counterfactual is not actionable if it alters immutable attributes, violates causal dependencies, or shifts costs to the affected person. Recourse must therefore distinguish predictive contrast from feasible intervention [35-37].

4.2.5 Knowledge-graph-based explanations

TransE evaluates knowledge-graph triple (h, r, t) using translational entity embeddings, with the scoring function:

$$f_r(h,t) = -\left\|e_h + e_r - e_t\right\|_2.$$

Here, e_h and e_t are the head and tail entity embeddings, respectively, and e_r is the relation embedding. Higher scores indicate greater compatibility. Paths and subgraphs can support semantic explanations, yet embedding proximity alone does not constitute a human-readable explanation. Instead, relevance is determined by relation quality, ontology coverage, and path selection [38].

4.2.6 Gradient, propagation, and visual explanations

Integrated Gradients (IG) attributes a model's prediction relative to a baseline x' for an input x along a straight-line path:

$$IG_i(x) = \left(x_i - x_i'\right)\int_0^1 \frac{\partial F\left(x' + \alpha\left(x - x'\right)\right)}{\partial x_i} d\alpha,$$

where F denotes the model output, x' represents the baseline, α serves as the path interpolation parameter, and i indexes the input coordinate. However, such attributions heavily depend on the choice of baseline and path [39]. Similarly, while gradient-weighted class activation mapping (Grad-CAM) localizes class-relevant convolutional regions [3], visual plausibility should not be confused with model faithfulness. Indeed, randomization and parameter-sensitivity tests demonstrate that some visually compelling saliency maps exhibit surprisingly weak dependency on the learned model weights [21].

4.2.7 Concept-based explanations

Concept activation vectors and concept bottleneck models explain predictions using human-named concepts rather than raw features [40-41]. Their relevance is particularly high when domain experts reason in terms of concepts, but validity depends on concept definition, annotation quality, completeness, and stability across populations. A concept layer can still conceal measurement bias or omit important latent factors.

4.3 Critical comparison under PDR

No method dominates all PDR dimensions. SHAP trades axiomatic consistency against dependence-sensitive semantics; LIME trades model independence against sampling instability; counterfactuals trade actionability against causal infeasibility; concept methods trade human relevance against concept validity; saliency maps trade visual accessibility against uncertain faithfulness; intrinsic models trade auditability against possible loss of flexibility in complex unstructured tasks; and post-hoc methods

can supply practical evidence while also rationalizing an opaque predictor. Adversarial constructions can deliberately produce reassuring LIME or SHAP explanations for biased models [42]. These tensions are summarized in the method-selection matrix in Table S3 of the Supplementary Material.

5. XAI applications across disciplines

Applications are compared using a common decision template: decision problem, data and predictor, explanation family, target user, PDR priority, evaluation evidence, and practical value. This structure distinguishes the presence of an XAI output from evidence that it improves a decision.

5.1 Environmental science

Environmental decisions combine spatiotemporal prediction with physical plausibility, public accountability, and transfer across locations. XAI has been used to interpret green-energy labor markets and energy-system models [43-44]. Attribution may identify drivers of load, pollution, or ecological risk, but an explanation that changes under spatial resampling or conflicts with known mechanisms should not guide policy. The dominant PDR priorities are predictive reliability and relevance; scientists and policymakers generally need stable global effects, uncertainty-aware SHAP/GAM summaries, and spatial diagnostics because decisions must transfer beyond a single fitted sample.

5.2 Education

Educational XAI supports early warning, feedback, personalization, and institutional intervention [45-46]. Students need actionable feedback, teachers need diagnostic evidence, and administrators need fairness and subgroup monitoring. Feature importance can stigmatize learners when correlations are treated as causes, and explanations may expose sensitive behavioral data. Relevance is therefore dominant; teachers and learners generally need sparse rule/GAM summaries and constrained counterfactual feedback because explanations must lead to pedagogically feasible actions.

5.3 Cybersecurity

Cybersecurity applications include malware, intrusion, anomaly, and threat detection [47-49]. Explanations must reduce investigation time, prioritize alerts, expose false positives, and resist adversarial manipulation; a static explanation can become obsolete as attackers adapt. Predictive accuracy and analyst relevance are co-dominant; security analysts generally need concise local attributions plus rule or prototype summaries because rapid triage must be paired with stability and adversarial stress testing.

5.4 Finance

Finance uses XAI in credit, fraud, portfolio, and risk decisions [50-52]. The same outputs often must support model validation, regulatory review, applicant communication, and recourse. SHAP can provide consistent additive summaries, but correlated variables and background choices can change their meaning; counterfactuals are useful only when actions are feasible and lawful. Descriptive accuracy and relevance are paramount; auditors and applicants generally need monotonic scorecards or GAMs, stable SHAP analyses with dependence checks, and constrained recourse because decisions must be contestable and auditable.

5.5 Law

In legal settings, XAI supports contestability, procedural fairness, evidentiary scrutiny, and allocation of responsibility [53-55]. A technically faithful feature weight is not automatically a legally sufficient reason; explanations must connect model evidence to admissible factors, institutional procedures, and affected rights. Relevance and descriptive accuracy are paramount; judges, lawyers, regulators, and affected parties generally need traceable rule paths, source-grounded evidence, and constrained local explanations because reasons must support challenge rather than merely persuasion.

5.6 Agriculture

Agricultural XAI supports yield, irrigation, disease, resource-allocation, and sustainability decisions under variable soil, climate, and management conditions [56]. Global feature effects can reveal agronomic drivers, image explanations can localize disease evidence, and counterfactuals can suggest management changes, but transfer across farms and seasons is a central limitation. Relevance and predictive reliability are dominant; farmers and extension agents generally need compact rules or GAM effects, localized visual evidence, and calibrated uncertainty because recommendations must be feasible under local conditions.

5.7 Healthcare

Healthcare XAI spans diagnosis, triage, prognosis, treatment, imaging, and patient communication [57-59]. SHAP and LIME can clarify tabular risk, Grad-CAM can localize imaging evidence, and counterfactuals can frame possible interventions, but none proves clinical validity. Explanations require external validation, uncertainty characterization, subgroup evidence, and workflow testing; otherwise, they can create false reassurance [60-61]. Predictive reliability and relevance are dominant; clinicians and patients generally need calibrated intrinsic models where feasible, clinically validated attribution or

localization, and uncertainty-aware summaries because explanations must support professional judgment rather than replace it.

The complete cross-domain comparison is provided in Table S4 in the Supplementary Material.

6. Emerging frontiers in decision-oriented XAI

6.1 Causal, uncertainty-aware, and robust explanations

Counterfactual explanations become decision-relevant only when predictive contrast is connected to a feasible intervention. Causal recourse must model immutable variables, dependencies among actions, uncertainty in structural assumptions, and the distribution of burden across groups [37]. Robustness requires testing explanation sensitivity across perturbations, random seeds, model specifications, and distributional shifts. Uncertainty-aware XAI should report confidence in both prediction and explanatory evidence rather than displaying a single deterministic ranking.

6.2 Concept-based explanations

Concept-based methods can align explanations with clinical signs, legal categories, object parts, or scientific constructs [40-41]. The research frontier is not simply discovering named concepts; rather, it is validating whether concepts are complete, causally meaningful, stable across populations, and actionable under intervention. Concept completeness and leakage tests should accompany human ratings.

6.3 Generative and multimodal AI

Multimodal systems combine text, images, signals, tables, and retrieved knowledge. Explanations must identify which modality provides decisive evidence, whether cross-modal evidence is consistent, and whether a generated rationale is grounded in the input. Additionally, generative-AI explainability raises verifiability, interaction, security, and computational-cost requirements [11]. PDR separates output quality from faithfulness and from the user's ability to verify or contest the output.

6.4 Large language models and foundation models

Foundation-model behavior reflects an interplay of pretraining data, instruction tuning, retrieval-augmented generation, tools, and context. A fluent chain-of-thought rationale may serve as effective communication while remaining an unfaithful account of the underlying mechanism producing the answer [10, 62]. Consequently, high-stakes applications should pair rationales with source provenance, counterfactual behavior tests, uncertainty communication, retrieval audits, and task-specific human validation. Crucially, mechanistic evidence, behavioral evidence, and user-facing explanation should be treated and reported as different objects.

7. Conclusion and future work

7.1 Conclusion

This LDA-assisted structured review mapped 666 verified OpenAlex records into eight empirically derived themes and interpreted them through an adopted PDR lens. The synthesis shows that explanation quality is conditional: predictive reliability determines whether the model deserves attention, descriptive accuracy determines whether an explanation reflects model behavior, and relevance determines whether the evidence helps a defined stakeholder decide, contest, audit, or act.

Across domains, method preference follows decision structure rather than popularity. Intrinsic models are especially valuable when global auditability and structured data make them competitive, while post-hoc methods are often necessary for high-dimensional predictors but require explicit fidelity and stability tests. The central practical lesson is to choose the explanation only after defining the user, action, risk, and evidence requirements.

7.2 Limitations and future-research roadmap

This review has bounded limitations. OpenAlex coverage differs from proprietary and domain-specific databases; abstracts and keywords are incomplete for some records; English-language screening introduces language bias; and the 2018-2026 window excludes earlier non-seminal work. LDA relies on bag-of-words assumptions and is sensitive to preprocessing choices, vocabulary thresholds, random seeds, and k. Topic labeling is interpretive, and metadata-based mapping cannot replace full-text synthesis. The rapidly evolving XAI literature means that foundation-model coverage will require periodic updates.

Future work should shift focus from explanation availability to decision evidence. Priorities include standardized PDR-aligned evaluation; context-dependent tests of prediction and interpretability; human-centered task validation; causal and actionable recourse; robustness, uncertainty, drift, and lifecycle monitoring; defenses against explanation gaming; privacy–transparency trade-offs; regulatory and legal accountability; protection of vulnerable stakeholders; multimodal and generative-AI faithfulness; computational and carbon costs; and reproducible deployment governance [19, 42, 63]. Table S5 in the Supplementary Material translates these frontiers into testable questions, recommended evaluation strategies,

expected decision value, and responsible stakeholders.

Supplementary information

Supplementary materials associated with this article are available online and can be accessed via the journal website: https://ojs.luminescience.cn/DMA/article/view/590/499.

Authors' contribution

Jiangshan Zhu: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Data curation, Visualization, Writing—original draft, and Writing—review and editing.

Funding

This research received no external funding.

Conflicts of interest

The author declares no conflicts of interest.

References

[1] Goodfellow I, Bengio Y, Courville A. *Deep Learning*. Cambridge (MA): MIT Press; 2016.

[2] Ali S, Abuhmed T, El-Sappagh S, Muhammad K, Alonso-Moral JM, Confalonieri R, et al. Explainable Artificial Intelligence (XAI): What we know and what is left to attain Trustworthy Artificial Intelligence. *Information Fusion*. 2023; 99: 101805. doi:10.1016/j.inffus.2023.101805.

[3] Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-CAM: Visual explanations from deep networks via gradient-based localization. *International Journal of Computer Vision*. 2020; 128(2): 336-359. doi:10.1007/s11263-019-01228-7.

[4] Miller T. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*. 2019; 267: 1-38. doi:10.1016/j.artint.2018.07.007.

[5] Murdoch WJ, Singh C, Kumbier K, Abbasi-Asl R, Yu B. Definitions, methods, and applications in interpretable machine learning. *Proceedings of the National Academy of Sciences*. 2019; 116(44): 22071-22080. doi:10.1073/pnas.1900654116.

[6] Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*. 2019; 1(5): 206-215. doi:10.1038/s42256-019-0048-x.

[7] Adadi A, Berrada M. Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). *IEEE Access*. 2018; 6: 52138-52160. doi:10.1109/ACCESS.2018.2870052.

[8] Gunning D, Vorm E, Wang JY, Turek M. DARPA's explainable AI (XAI) program: A retrospective. *Applied AI Letters*. 2021; 2(4): e61. doi:10.1002/ail2.61.

[9] Saeed W, Omlin C. Explainable AI (XAI): A systematic meta-survey of current challenges and future opportunities. *Knowledge-Based Systems*. 2023; 263: 110273. doi:10.1016/j.knosys.2023.110273.

[10] Zhao H, Chen H, Yang F, Liu N, Deng H, Cai H, et al. Explainability for large language models: A survey. *ACM Transactions on Intelligent Systems and Technology*. 2024; 15(2): 1-38. doi:10.1145/3639372.

[11] Schneider J. Explainable Generative AI (GenXAI): A survey, conceptualization, and research agenda. *Artificial Intelligence Review*. 2024; 57(11): 289. doi:10.1007/s10462-024-10916-x.

[12] Blei DM, Ng AY, Jordan MI. Latent Dirichlet allocation. *Journal of Machine Learning Research.* 2003; 3: 993-1022.

[13] Griffiths TL, Steyvers M. Finding scientific topics. *Proceedings of the National Academy of Sciences*. 2004; 101(suppl_1): 5228-5235. doi:10.1073/pnas.0307752101.

[14] Mimno D, Wallach H, Talley E, Leenders M, McCallum A. Optimizing semantic coherence in topic models. In: *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*. USA: Association for Computational Linguistics; 2011. p.262-272.

[15] Newman D, Lau JH, Grieser K, Baldwin T. Automatic evaluation of topic coherence. In: *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. USA: Association for Computational Linguistics; 2010. p.100-108.

[16] Roder M, Both A, Hinneburg A. Exploring the space of topic coherence measures. In: *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*. New York, USA: Association for Computing Machinery; 2015. p.399-408. doi:10.1145/2684822.2685324.

[17] Cao J, Xia T, Li J, Zhang Y, Tang S. A density-based method for adaptive LDA model selection. *Neurocomputing*. 2009; 72(7-9): 1775-1781. doi:10.1016/j.neucom.2008.06.011.

[18] Sievert C, Shirley K. LDAvis: A method for visualizing and interpreting topics. In: *Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces*. USA: Association for Computational Linguistics. 2014. p.63-70. doi:10.3115/v1/W14-3110.

[19] Nauta M, Trienes J, Pathak S, Nguyen E, Peters M, Schmitt Y, et al. From anecdotal evidence to quantitative evaluation methods: A systematic review

on evaluating explainable AI. *ACM Computing Surveys*. 2023; 55(13s): 1-42. doi:10.1145/3583558.

[20] Yeh CK, Hsieh CY, Suggala A, Inouye DI, Ravikumar PK. On the (in)fidelity and sensitivity of explanations. In: *33rd Conference on Neural Information Processing Systems (NeurIPS 2019)*. USA: Neural Information Processing Systems Foundation; 2019. p.10965-10976.

[21] Adebayo J, Gilmer J, Muelly M, Goodfellow I, Hardt M, Kim B. Sanity checks for saliency maps. In: *Advances in Neural Information Processing Systems 31*. USA: Curran Associates, Inc.; 2018. p.9505-9515.

[22] Lipton ZC. The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue*. 2018; 16(3): 31-57. doi:10.1145/3236386.3241340.

[23] Lou Y, Caruana R, Gehrke J, Hooker G. Accurate intelligible models with pairwise interactions. In: *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York: Association for Computing Machinery; 2013. p.623-631. doi:10.1145/2487575.2487579.

[24] Ravikumar P, Lafferty J, Liu H, Wasserman L. Sparse additive models. *Journal of the Royal Statistical Society Series B: Statistical Methodology*. 2009; 71(5): 1009-1030. doi:10.1111/j.1467-9868.2009.00718.x.

[25] Breiman L, Friedman JH, Olshen RA, Stone CJ. *Classification and Regression Trees*. New York: Chapman and Hall/CRC; 1984.

[26] Sagi O, Rokach L. Explainable decision forest: Transforming a decision forest into an interpretable tree. *Information Fusion*. 2020; 61: 124-138. doi:10.1016/j.inffus.2020.03.013.

[27] Letham B, Rudin C, McCormick TH, Madigan D. Interpretable classifiers using rules and bayesian analysis: Building a better stroke prediction model. *Annals of Applied Statistics*. 2015; 9(3): 1350-1371. doi:10.1214/15-AOAS848.

[28] Cohen WW. Fast effective rule induction. In: *Machine Learning Proceedings 1995*. USA: Morgan Kaufmann Publishers; 1995. p.115-123. doi:10.1016/b978-1-55860-377-6.50023-2.

[29] Clark P, Niblett T. The CN2 induction algorithm. *Machine Learning*. 1989; 3(4): 261-283. doi:10.1007/BF00116835.

[30] Lakkaraju H, Bach SH, Leskovec J. Interpretable decision sets: A joint framework for description and prediction. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York: Association for Computing Machinery; 2016. p.1675-1684. doi:10.1145/2939672.2939874.

[31] Kim B, Khanna R, Koyejo OO. Examples are not enough, learn to criticize! Criticism for interpretability. In: *Advances in Neural Information Processing Systems 29*. Red Hook, NY: Curran Associates; 2016. p.2288-2296.

[32] Koh PW, Liang P. Understanding black-box predictions via influence functions. In: *Proceedings of the 34th International Conference on Machine Learning*. Sydney, Australia: Proceedings of Machine Learning Research; 2017. p.1885-1894.

[33] Lundberg SM, Lee SI. A unified approach to interpreting model predictions. In: *Advances in Neural Information Processing Systems 30*. Red Hook, NY: Curran Associates; 2017. p.4765-4774.

[34] Ribeiro MT, Singh S, Guestrin C. "Why should I trust you?" Explaining the predictions of any classifier. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York: Association for Computing Machinery; 2016. p.1135-1144. doi:10.1145/2939672.2939778.

[35] Wachter S, Mittelstadt B, Russell C. Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harvard Journal of Law & Technology*. 2017; 31(2): 841-887.

[36] Mothilal RK, Sharma A, Tan C. Explaining machine learning classifiers through diverse counterfactual explanations. In: *Proceedings of the Conference on Fairness, Accountability, and Transparency*. New York: Association for Computing Machinery; 2020. p.607-617. doi:10.1145/3351095.3372850.

[37] Karimi AH, Scholkopf B, Valera I. Algorithmic recourse: From counterfactual explanations to interventions. In: *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. New York: Association for Computing Machinery; 2021. p.353-362. doi:10.1145/3442188.3445899.

[38] Bordes A, Usunier N, Garcia-Duran A, Weston J, Yakhnenko O. Translating embeddings for modeling multi-relational data. In: *Advances in Neural Information Processing Systems 26*. Red Hook, NY: Curran Associates; 2013. p.2787-2795.

[39] Sundararajan M, Taly A, Yan Q. Axiomatic attribution for deep networks. In: *Proceedings of the 34th International Conference on Machine Learning*. Sydney, Australia: Proceedings of Machine Learning Research; 2017. p.3319-3328.

[40] Kim B, Wattenberg M, Gilmer J, Cai C, Wexler J, Viegas F, et al. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV). In: *Proceedings of the 35th International Conference on Machine Learning*. Stockholm, Sweden: Proceedings of Machine Learning Research; 2018. p.2668-2677.

[41] Koh PW, Nguyen T, Tang YS, Mussmann S, Pierson E, Kim B, et al. Concept bottleneck models. In: *Proceedings of the 37th International Conference on Machine Learning*. Virtual Event: Proceedings of Machine Learning Research; 2020. p.5338-5348.

[42] Slack D, Hilgard S, Jia E, Singh S, Lakkaraju H. Fooling lime and shap: Adversarial attacks on post hoc explanation methods. In: *Proceedings of the*

AAAI/ACM Conference on AI, Ethics, and Society. New York: Association for Computing Machinery; 2020. p.180-186. doi:10.1145/3375627.3375830.

[43] Chen H, Mason CM. Explainable AI (XAI) for constructing a lexicon for classifying green energy jobs: A comparative analysis of occupation, industry, and location composition with traditional energy jobs. *IEEE Access*. 2024; 12: 142709-142720. doi:10.1109/ACCESS.2024.3430317.

[44] Machlev R, Heistrene L, Perl M, Levy KY, Belikov J, Mannor S, et al. Explainable Artificial Intelligence (XAI) techniques for energy and power systems: Review, challenges and opportunities. *Energy and AI*. 2022; 9: 100169. doi:10.1016/j.egyai.2022.100169.

[45] Khosravi H, Shum SB, Chen G, Conati C, Tsai YS, Kay J, et al. Explainable Artificial Intelligence in education. *Computers and Education: Artificial Intelligence*. 2022; 3: 100074. doi:10.1016/j.caeai.2022.100074.

[46] Rachha A, Seyam M. Explainable AI in education: Current trends, challenges, and opportunities. In: *SoutheastCon 2023*. Orlando, FL, USA: IEEE; 2023. p.232-239. doi:10.1109/SOUTHEASTCON51012.2023.10115140.

[47] Rjoub G, Bentahar J, Abdel Wahab O, Mizouni R, Song A, Cohen R, et al. A survey on explainable artificial intelligence for cybersecurity. *IEEE Transactions on Network and Service Management*. 2023; 20(4): 5115-5140. doi:10.1109/TNSM.2023.3282740.

[48] Hariharan S, Robinson RRR, Prasad RR, Thomas C, Balakrishnan N. XAI for intrusion detection system: Comparing explanations based on global and local scope. *Journal of Computer Virology and Hacking Techniques*. 2023; 19(2): 217-239. doi:10.1007/s11416-022-00441-2.

[49] Sarcevic A, Pintar D, Vranic M, Krajna A. Cybersecurity knowledge extraction using XAI. *Applied Sciences*. 2022; 12(17): 8669. doi:10.3390/app12178669.

[50] Weber P, Carl KV, Hinz O. Applications of explainable artificial intelligence in finance—a systematic review of finance, information systems, and computer science literature. *Management Review Quarterly*. 2024; 74(2): 867-907. doi:10.1007/s11301-023-00320-0.

[51] Awosika T, Shukla RM, Pranggono B. Transparency and privacy: The role of explainable AI and federated learning in financial fraud detection. *IEEE Access*. 2024; 12: 64551-64560. doi:10.1109/ACCESS.2024.3394528.

[52] Bussmann N, Giudici P, Marinelli D, Papenbrock J. Explainable AI in fintech risk management. *Frontiers in Artificial Intelligence*. 2020; 3: 26. doi:10.3389/frai.2020.00026.

[53] Hacker P, Krestel R, Grundmann S, Naumann F. Explainable AI under contract and tort law: Legal incentives and technical challenges. *Artificial Intelligence and Law*. 2020; 28(4): 415-439. doi:10.1007/s10506-020-09260-6.

[54] Richmond KM, Muddamsetty SM, Gammeltoft-Hansen T, Olsen HP, Moeslund TB. Explainable AI and law: An evidential survey. *Digital Society*. 2024; 3(1): 1. doi:10.1007/s44206-023-00081-z.

[55] Vale D, El-Sharif A, Ali M. Explainable artificial intelligence (XAI) post-hoc explainability methods: Risks and limitations in non-discrimination law. *AI and Ethics*. 2022; 2(4): 815-826. doi:10.1007/s43681-022-00142-y.

[56] Ryo M. Explainable artificial intelligence and interpretable machine learning for agricultural data analysis. *Artificial Intelligence in Agriculture*. 2022; 6: 257-265. doi:10.1016/j.aiia.2022.11.003.

[57] Saraswat D, Bhattacharya P, Verma A, Prasad VK, Tanwar S, Sharma G, et al. Explainable AI for healthcare 5.0: Opportunities and challenges. *IEEE Access*. 2022; 10: 84486-84517. doi:10.1109/ACCESS.2022.3197671.

[58] Sadeghi Z, Alizadehsani R, Cifci MA, Kausar S, Rehman R, Mahanta P, et al. A review of explainable artificial intelligence in healthcare. *Computers and Electrical Engineering*. 2024; 118: 109370. doi:10.1016/j.compeleceng.2024.109370.

[59] Okada Y, Ning Y, Ong MEH. Explainable artificial intelligence in emergency medicine: An overview. *Clinical and Experimental Emergency Medicine*. 2023; 10(4): 354-362. doi:10.15441/ceem.23.145.

[60] Ghassemi M, Oakden-Rayner L, Beam AL. The false hope of current approaches to explainable artificial intelligence in health care. *The Lancet Digital Health*. 2021; 3(11): e745-e750. doi:10.1016/S2589-7500(21)00208-9.

[61] Ahmed S, Kaiser MS, Hossain MS, Andersson K. A comparative analysis of LIME and SHAP interpreters with explainable ML-based diabetes predictions. *IEEE Access*. 2025; 13: 37370-37388. doi:10.1109/ACCESS.2024.3422319.

[62] Turpin M, Michael J, Perez E, Bowman S. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. In: *37th Conference on Neural Information Processing Systems*. Red Hook, NY: Curran Associates; 2023. p.1-14.

[63] Raji ID, Smart A, White RN, Mitchell M, Gebru T, Hutchinson B, et al. Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. In: *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. New York: Association for Computing Machinery; 2020. p.33-44. doi:10.1145/3351095.3372873.

[64] Wang Z, Samsten I, Miliou I, Mochaourab R, Papapetrou P. Glacier: Guided locally constrained counterfactual explanations for time series classification. *Machine Learning*. 2024; 113(7): 4639-4669. doi:10.1007/s10994-023-06502-x.

[65] Prenkaj B, Villaizán-Vallelado M, Leemann T, Kasneci G. Adapting to change: Robust counterfactual explanations in dynamic data landscapes. In: Meo R, Silvestri F. (eds.) *Machine Learning and Principles and Practice of Knowledge Discovery in Databases*. Cham: Springer; 2025. p.1-15. doi:10.1007/978-3-031-74630-7_22.

[66] Jiang J, Leofante F, Rago A, Toni F. Formalising the robustness of counterfactual explanations for neural networks. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Washington, DC: AAAI Press; 2023. p.14901-14909. doi:10.1609/aaai.v37i12.26740.

[67] Suresh H, Gomez SR, Nam KK, Satyanarayan A. Beyond expertise and roles: A framework to characterize the stakeholders of interpretable machine learning and their needs. In: *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. New York: Association for Computing Machinery; 2021. p.1-16. doi:10.1145/3411764.3445088.

[68] Leichtmann B, Humer C, Hinterreiter A, Streit M, Mara M. Effects of explainable artificial intelligence on trust and human behavior in a high-risk decision task. *Computers in Human Behavior*. 2023; 139: 107539. doi:10.1016/j.chb.2022.107539.

[69] Nannini L. Habemus a right to an explanation: So what? A framework on transparency-explainability functionality and tensions in the EU AI Act. In: *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*. New York: Association for Computing Machinery; 2024; 7(1): 1023-1035. doi:10.1609/aies.v7i1.31700.

[70] Zhang H, Yang YF, Song XL, Hu HJ, Yang YY, Zhu X, et al. An interpretable artificial intelligence model based on CT for prognosis of intracerebral hemorrhage: A multicenter study. *BMC Medical Imaging*. 2024; 24: 170. doi:10.1186/s12880-024-01352-y.

[71] Torres-Martos Á, Anguita-Ruiz A, Bustos-Aibar M, Ramírez-Mena A, Arteaga M, Bueno G, et al. Multiomics and explainable artificial intelligence for decision support in insulin resistance early diagnosis: A pediatric population-based longitudinal study. *Artificial Intelligence in Medicine*. 2024; 156: 102962. doi:10.1016/j.artmed.2024.102962.

[72] Casiraghi E, Malchiodi D, Trucco G, Frasca M, Cappelletti L, Fontana T, et al. Explainable machine learning for early assessment of COVID-19 risk prediction in emergency departments. *IEEE Access*. 2020; 8: 196299-196325. doi:10.1109/ACCESS.2020.3034032.

[73] Moosavi S, Farajzadeh-Zanjani M, Razavi-Far R, Palade V, Saif M. Explainable AI in manufacturing and industrial cyber–physical systems: A survey. *Electronics*. 2024; 13(17): 3497. doi:10.3390/electronics13173497.

[74] Capuano N, Fenza G, Loia V, Stanzione C. Explainable artificial intelligence in cybersecurity: A survey. *IEEE Access*. 2022; 10: 93575-93600. doi:10.1109/ACCESS.2022.3204171.

[75] Neupane S, Ables J, Anderson W, Mittal S, Rahimi S, Banicescu I, et al. Explainable intrusion detection systems (X-IDS): A survey of current methods, challenges, and opportunities. *IEEE Access*. 2022; 10: 112392-112415. doi:10.1109/ACCESS.2022.3216617.

[76] Guidotti R. Counterfactual explanations and how to find them: Literature review and benchmarking. *Data Mining and Knowledge Discovery*. 2024; 38(5): 2770-2824. doi:10.1007/s10618-022-00831-6.

[77] De Toni G, Lepri B, Passerini A. Synthesizing explainable counterfactual policies for algorithmic recourse with program synthesis. *Machine Learning*. 2023; 112(4): 1389-1409. doi:10.1007/s10994-022-06293-7.

[78] Lamy JB, Sekar B, Guezennec G, Bouaud J, Seroussi B. Explainable artificial intelligence for breast cancer: A visual case-based reasoning approach. *Artificial Intelligence in Medicine*. 2019; 94: 42-53. doi:10.1016/j.artmed.2019.01.001.

[79] Patrício C, Neves JC, Teixeira LF. Explainable deep learning methods in medical image classification: A survey. *ACM Computing Surveys*. 2024; 56(4): 1-41. doi:10.1145/3625287.

[80] Abdullakutty F, Akbari Y, Al-Maadeed S, Bouridane A, Talaat IM, Hamoudi R. Histopathology in focus: A review on explainable multi-modal approaches for breast cancer diagnosis. *Frontiers in Medicine*. 2024; 11: 1450103. doi:10.3389/fmed.2024.1450103.

[81] Rosenbacke R, Melhus Å, McKee M, Stuckler D. How explainable artificial intelligence can increase or decrease clinicians' trust in AI applications in health care: Systematic review. *JMIR AI*. 2024; 3: e53207. doi:10.2196/53207.

[82] Pierce RL, Van Biesen W, Van Cauwenberge D, Decruyenaere J, Sterckx S. Explainability in medicine in an era of AI-based clinical decision support systems. *Frontiers in Genetics*. 2022; 13: 903600. doi:10.3389/fgene.2022.903600.

[83] Rong Y, Leemann T, Nguyen TT, Fiedler L, Qian P, Unhelkar V, et al. Towards human-centered explainable AI: A survey of user studies for model explanations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 2024; 46(4): 2104-2122. doi:10.1109/TPAMI.2023.3331846.

[84] Kim J, Maathuis H, Sent D. Human-centered evaluation of explainable AI applications: A systematic review. *Frontiers in Artificial Intelligence*. 2024; 7: 1456486. doi:10.3389/frai.2024.1456486.

[85] Senoner J, Schallmoser S, Kratzwald B, Feuerriegel S, Netland T. Explainable AI improves task performance in human-AI collaboration. *Scientific Reports*. 2024; 14(1): 31150. doi:10.1038/s41598-024-82501-9.