Supplementary Material

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

Manuscript ID: DMA-590

Table S1. Comparison of the present review with representative XAI surveys

Study	Review design/data basis	Primary organizing logic	Method-comparison depth	Human/decision orientation	Empirical literature mapping	LLM/foundation-model coverage	Distinct contribution or limitation
Adadi and Berrada [7]	Narrative survey	XAI concepts and taxonomy	Broad method overview	Trust and user needs discussed	No corpus model	No	Foundational field map; predates current foundation models
Murdoch et al. [5]	Conceptual review	PDR and interpretable ML	Selective, principle-led	Audience relevance is explicit	No corpus model	No	Supplies the PDR basis adopted here
Gunning et al. [8]	Program retrospective	DARPA XAI goals and projects	Program-level synthesis	Human understanding central	Program evidence	No	Explains program trajectory rather than surveying the whole field

Study	Review design/data basis	Primary organizing logic	Method-comparison depth	Human/decision orientation	Empirical literature mapping	LLM/foundation-model coverage	Distinct contribution or limitation
Ali et al. [2]	Comprehensive review	Trustworthy-AI requirements	Extensive taxonomy	Stakeholders and trust covered	Narrative mapping	Limited	Broad contemporary synthesis
Saeed and Omlin [9]	Systematic meta-survey	Challenges and opportunities	Compares survey findings	Human factors included	Survey-level synthesis	Limited	Meta-level consolidation of XAI challenges
Zhao et al. [10]	Focused survey	LLM explanation methods	Detailed for LLMs	Use cases considered	Narrative mapping	Yes	Deep LLM coverage; not cross-domain decision synthesis
Schneider [11]	Narrative/integrative review	Generative-AI explainability	GenAI-specific taxonomy	Verifiability and interaction emphasized	Qualitative mapping	Yes	Focused roadmap for generative AI
Present review	LDA-assisted structured review (n=666)	PDR-based decision orientation	Critical method comparison	Stakeholder, actionability, and risk explicit	LDA topics plus manual interpretation	Yes	Integrates empirical mapping, method trade-offs, application matrix, and roadmap

Table S2. LDA topic clustering and manual coding results

Topic and prevalence	First-order LDA topic	Representative keywords	Verified representative papers	Manual interpretation	Primary PDR connection	Main research gap
T1: Robust counterfactual generation (13.1%)	Counterfactual explanations and algorithmic recourse	counterfactual_explanation; robustness; optimization; time series; recourse; action; state; input; feasible set	Wang et al. [64]; Prenkaj et al. [65]; Jiang et al. [66]	Optimization-centered methods generate sparse and robust contrasts for dynamic, temporal, or constrained prediction settings.	Descriptive accuracy plus relevance: the contrast must be faithful, stable, and feasible.	Robustness under distribution shift, temporal change, and uncertain feasibility remains weakly standardized.
T2: Responsible and regulated decision support (9.9%)	Human-centered trust, fairness, and governance	transparency; fairness; stakeholder; bias; trust; accountability; finance; credit; policy; human	Suresh et al. [67]; Leichtmann et al. [68]; Nannini [69]	Stakeholder needs, trust calibration, fairness, and regulatory accountability shape explanations in financial and public decisions.	Relevance: reasons must fit stakeholder rights, oversight duties, and decision risk.	Legal sufficiency, calibrated reliance, and equitable explanation burdens need outcome-based evaluation.

Topic and prevalence	First-order LDA topic	Representative keywords	Verified representative papers	Manual interpretation	Primary PDR connection	Main research gap
T3: Interpretable clinical risk prediction (16.5%)	Feature attribution and surrogate explanation in predictive models	disease; SHAP; patient; risk; clinical; diagnosis; random forest; ensemble; early detection; Shapley	Zhang et al. [70]; Torres-Martos et al. [71]; Casiraghi et al. [72]	Clinical prediction studies combine tree ensembles and feature attribution to identify patient-level risk drivers.	Predictive and descriptive accuracy: calibrated risk must be paired with stable case-level evidence.	External validation, subgroup calibration, attribution stability, and demonstrated workflow benefit remain uneven.
T4: Industrial, cybersecurity, and infrastructure XAI (11.6%)	Security, fraud, and risk monitoring	security; detection; energy; cybersecurity; industry; network; intrusion; attack; SHAP; LIME	Moosavi et al. [73]; Capuano et al. [74]; Neupane et al. [75]	Operational XAI links anomaly detection, cyber defense, predictive maintenance, and infrastructure decisions.	Predictive accuracy plus relevance: explanations must accelerate triage and remain attack resistant.	Real-time scalability, adversarial robustness, privacy-preserving audit, and explanation drift need joint tests.

Topic and prevalence	First-order LDA topic	Representative keywords	Verified representative papers	Manual interpretation	Primary PDR connection	Main research gap
T5: Causal and actionable algorithmic recourse (11.2%)	Counterfactual explanations and algorithmic recourse	counterfactual_explanation; causal; recourse; outcome; individual; algorithmic_recourse; action; policy; agent; economics	Karimi et al. [37]; Guidotti [76]; De Toni et al. [77]	Recourse research moves from individual prediction contrasts toward causal interventions, summaries, and policies.	Relevance: a formally valid contrast must become a feasible and equitable intervention.	Causal assumptions, long-term validity, group burden, and recourse gaming require stronger evidence.
T6: Visual and multimodal medical explanation (10.8%)	Deep, visual, and multimodal explanation	image; deep_learning; medical; cancer; neural network; diagnosis; convolutional; Grad-CAM; lesion; multimodal	Lamy et al. [78]; Patrício et al. [79]; Abdullakutty et al. [80]	Visual attribution, case-based reasoning, and multimodal evidence make imaging and radiomics predictions inspectable.	Descriptive accuracy: visual plausibility must be tested against model dependence and clinical localization.	Faithfulness, baseline sensitivity, uncertainty, multimodal consistency, and external-site validation remain open.

Topic and prevalence	First-order LDA topic	Representative keywords	Verified representative papers	Manual interpretation	Primary PDR connection	Main research gap
T7: Clinical XAI adoption, trust, and workflow (12.5%)	Clinical and health decision support	healthcare; clinical; medicine; trust; patient; decision_support; clinician; transparency; adoption; care	Saraswat et al. [57]; Rosenbacke et al. [81]; Pierce et al. [82]	Clinical reviews examine whether explanation formats support safe adoption, professional judgment, and patient communication.	Relevance plus predictive reliability: evidence must fit clinical workflow and calibrated reliance.	Prospective task outcomes, automation bias controls, patient perspectives, and safety monitoring are limited.
T8: Human-centered XAI evaluation and taxonomy (14.3%)	Human-centered trust, fairness, and governance	user; human; design; taxonomy; language; interaction; expert; concept; metric; explanation	Rong et al. [83]; Kim et al. [84]; Senoner et al. [85]	Surveys and user studies organize explanation forms and evaluate comprehension, interaction, and collaborative performance.	Relevance: explanation quality is demonstrated through stakeholder task outcomes, not plausibility alone.	Comparable user-task protocols, longitudinal reliance, accessibility, and construct-validity metrics are needed.

Table S3. Comparison of representative XAI methods under the PDR framework

Method family	Type, scope, and dependence	Core assumptions	Computational complexity/scalability	Fidelity and stability	Actionability	Strengths and limitations	Typical decision contexts
Linear/logistic models	Intrinsic; global/local contributions; model specific	Correct link, specified effects, manageable dependence	Low; highly scalable	Exact for fitted structure; unstable under collinearity	Low unless action variables are explicit	Auditable; may miss nonlinearities and interactions	Regulated tabular scoring and policy models
GAM/GA2M/SpAM	Intrinsic; global; model specific	Additivity plus selected interactions and smoothness	Low-moderate; scalable with controls	Exact for fitted components; sensitive to smoothing/support	Moderate	Flexible effect curves; interaction proliferation can obscure	Clinical risk, pricing, environmental response
Trees/rule lists/rule sets	Intrinsic; global and path-local; model specific	Axis-aligned or symbolic decision structure	Low-moderate; pruning/search varies	Exact logic but structurally unstable	Moderate-high	Simulatable; overlap, conflict, and fragmentation	Eligibility, triage, audit, expert protocols
MMD-Critic/prototypes	Post-hoc; local/global examples; model agnostic	Meaningful similarity kernel and coverage	Moderate; kernel scaling can be costly	Data-representative, not necessarily model faithful	Low-moderate	Concrete examples; may omit rare but important cases	Case-based review and dataset auditing
Influence functions	Post-hoc; local; model specific	Differentiability and locally invertible Hessian	High for large models; approximation needed	Approximate and optimization dependent	Low	Traces training influence; fragile under non-convexity	Debugging, data quality, provenance

Method family	Type, scope, and dependence	Core assumptions	Computational complexity/scalability	Fidelity and stability	Actionability	Strengths and limitations	Typical decision contexts
SHAP	Post-hoc; local/global aggregation; broad model support	Defined coalition value and background distribution	Moderate-very high; approximations common	Axiomatic but semantics/background sensitive	Low-moderate	Comparable attributions; dependence and gaming risks	Audit, case explanation, feature monitoring
LIME	Post-hoc; local; model agnostic	Local smoothness and valid perturbation neighborhood	Moderate per case	Local fidelity measurable; sampling instability	Low	Simple surrogate; neighborhood can be unrealistic	Case review, debugging, user communication
Counterfactuals/recourse	Post-hoc; local; model agnostic or specific	Distance, feasibility, and causal constraints	Moderate-high; optimization dependent	Outcome contrast can be exact; model/data drift sensitive	High when feasible	Action-oriented; may recommend impossible or unfair changes	Credit, hiring, care planning, appeals
Knowledge graphs	Intrinsic/post-hoc; global/path-local; model specific	Valid ontology, entities, and relation semantics	Moderate-high; graph size dependent	Path fidelity depends on linkage and ranking	Moderate	Semantic traceability; incomplete or biased graph	Diagnosis, compliance, scientific reasoning

Method family	Type, scope, and dependence	Core assumptions	Computational complexity/scalability	Fidelity and stability	Actionability	Strengths and limitations	Typical decision contexts
IG/Grad-CAM/saliency	Post-hoc; local; model specific	Differentiability, baseline/layer choice	Low-moderate	Can fail randomization; baseline sensitive	Low	Modality-aligned visualization; plausibility can mislead	Imaging, signals, text diagnostics
Concept methods	Intrinsic/post-hoc; local/global; model specific	Valid, complete, stable concepts	Moderate-high; annotation intensive	Depends on concept layer and intervention tests	Moderate-high	Expert vocabulary; concept leakage and incompleteness	Clinical, legal, scientific, multimodal tasks

Table S4. Cross-domain decision-oriented XAI application matrix

Domain	Decision problem	Data/model family	Preferred explanation family	Target user	Dominant PDR priority and evaluation	Decision value and main limitation
Environmental science	Forecasting, monitoring, policy prioritization	Spatial-temporal/tabular; ensembles, deep learning (DL), generalized additive models (GAMs)	Stable global effects, uncertainty-aware SHAP, spatial diagnostics	Scientists, operators, policy-makers	Predictive reliability + relevance; spatial transfer, calibration, physical consistency	Mechanism checking and communication; limited by distribution shift
Education	Early warning, feedback, intervention	Learner records/text; trees, ensembles, sequence models	Rules/GAMs, constrained local attribution and counterfactual feedback	Students, teachers, administrators	Relevance; task outcomes, fairness, privacy, comprehension	Actionable support; risk of stigma and correlation-as-cause
Cybersecurity	Alert triage, intrusion/malware analysis	Logs/network/graphs; ensembles and DL	Local attribution plus rules/prototypes and adversarial tests	Security analysts, incident responders	Predictive accuracy + relevance; time-to-decision, stability, attack resistance	Faster investigation; explanations may be gamed
Finance	Credit, fraud, portfolio and risk decisions	Tabular/time series; scorecards, boosting, ensembles	Monotonic models, stable SHAP, feasible recourse	Applicants, auditors, model-risk teams, regulators	Descriptive accuracy + relevance; contestability, dependence tests, fairness	Audit and recourse; legal feasibility and correlated inputs matter

Domain	Decision problem	Data/model family	Preferred explanation family	Target user	Dominant PDR priority and evaluation	Decision value and main limitation
Law	Evidence review, eligibility, legal risk	Text/tabular/knowledge graphs	Rule paths, source-grounded evidence, constrained local explanations	Judges, lawyers, regulators, affected parties	Relevance + descriptive accuracy; procedural validity, traceability	Supports challenge; feature weights are not legal reasons
Agriculture	Yield, irrigation, disease and resource planning	Remote sensing, weather, soil; trees, DL, ensembles	GAM/rules, visual localization, calibrated uncertainty	Farmers, agronomists, extension agents	Relevance + predictive reliability; local transfer and feasibility	Actionable management; site and season dependence
Healthcare	Diagnosis, triage, prognosis, treatment	Electronic health records (EHR), imaging, signals; generalized additive models (GAMs), ensembles, deep learning (DL)	Calibrated intrinsic models, validated attribution/localization, uncertainty	Clinicians, patients, safety and regulatory teams	Predictive reliability + relevance; external validity, safety, workflow benefit	Clinical support; false reassurance and site shift remain risks

Table S5. Research roadmap for decision-oriented XAI

Research frontier	Current limitation	Core research question	Recommended methods/evaluation	Expected decision value	Main stakeholder
PDR-aligned standardized evaluation	Metrics are fragmented and often method specific	Which evidence jointly establishes prediction, fidelity, and relevance?	Shared benchmark tasks; calibration, fidelity, stability, and user-task protocols	Comparable claims and safer method selection	Researchers, standards bodies, regulators
Human-centered decision validation	Plausibility and satisfaction substitute for decision outcomes	When does an explanation improve decisions without inducing automation bias?	Pre-registered user studies; time, error, calibration, reliance, and harm measures	Workflow benefit and calibrated trust	End users, domain professionals, affected people
Causal and actionable recourse	Predictive counterfactuals can be infeasible or unfair	Which interventions are feasible, causal, stable, and equitably distributed?	Structural causal models; feasibility sets; recourse cost and subgroup audits	Actionable and contestable outcomes	Applicants, clinicians, case workers, regulators
Robustness, uncertainty, and explanation drift	Explanations vary across seeds, models, and shifts	When is explanatory evidence stable enough for consequential use?	Sensitivity tests; ensembles; uncertainty intervals; drift alarms; lifecycle monitoring	Reliable evidence over time	Model-risk teams, operators, auditors
Multimodal, generative, and large language model (LLM) explainability	Fluent rationales may be ungrounded or unfaithful	How can behavior, mechanism, source, and user-facing rationale be separated and verified?	Provenance tracing; modality ablation; causal/behavioral tests; faithfulness benchmarks	Verifiable AI-assisted decisions	Developers, users, safety teams, content owners

Research frontier	Current limitation	Core research question	Recommended methods/evaluation	Expected decision value	Main stakeholder
Governance, regulation, and ethical accountability	Transparency can conflict with privacy, security, and vulnerable-user protection	What explanation rights and audit evidence are proportionate to risk?	Algorithmic impact assessment; documentation; red-team testing; appeals; privacy-preserving audit	Accountability, due process, and harm prevention	Regulators, organizations, affected communities