Explainability Under Distribution Shift: Adaptive Explainable Artificial Intelligence for High-Risk Decision Systems

Patrick M Reynolds [1*], Vanessa B Cruz [2]
[1, 2] Trulaske College of Business, University of Missouri, Missouri, USA

* Corresponding Author: **Patrick M Reynolds**

Article Info

P-ISSN: 3051-3383
E-ISSN: 3051-3391
Volume: 06
Issue: 01
January – June 2025
Received: 14-12-2024
Accepted: 09-01-2025
Published: 25-01-2025
Page No: 13-24

Abstract
High-risk artificial intelligence systems are usually validated under a stable-data assumption, even though the distributions that shape predictions and explanations change after deployment. This study examines predictive degradation, explanation degradation, uncertainty, subgroup disparity, and adaptive governance across four temporally ordered decision settings: healthcare utilization, macroeconomic financial risk, energy-climate planning, and public-sector vulnerability prioritization. The empirical design combines real public data with natural temporal shifts in three domains and a documented controlled shift applied to real healthcare covariates. Logistic regression, random forest, and gradient boosting are evaluated in pre-shift, shift, and post-shift windows. Feature-ablation explanations are assessed through fidelity, temporal rank stability, directional consistency, actionability, subgroup disparity, and uncertainty. These components form a six-part Explanation Degradation Index. Static, recalibrated, periodically retrained, rolling, expanding, drift-triggered, and joint model-explanation strategies are compared under blocked temporal evaluation. Results show that predictive and explanation degradation are related but not interchangeable. Mean EDI increased by 0.032 during shift, while cross-domain changes in precision-recall performance were heterogeneous and statistically inconclusive. Recalibration frequently improved Brier score but left EDI unchanged. Joint adaptation with uncertainty-based deferral improved mean precision-recall area by 0.020 across eight domain-period tasks, although pairwise gains were not statistically decisive and finance exhibited a negative shift-period result. The Adaptive Explainability Under Shift Framework translates these findings into monitoring, escalation, adaptation, approval, rollback, and organizational-learning controls. The study contributes a measurable account of explanation reliability under change and a governance design that separates predictive confidence from explanation trustworthiness.

DOI: https://doi.org/10.54660/IJAIET.2025.6.1.13-24

Keywords: adaptive explainable artificial intelligence, distribution shift, concept drift, explanation stability, high-risk decision systems, responsible AI, model monitoring, human oversight

1. Introduction

High-risk decision systems are deployed into environments that do not remain fixed. Clinical practice changes as diagnostic capacity, treatment protocols, and patient populations change. Fraud systems encounter strategic adaptation, new payment channels, and revised investigation procedures. Energy forecasts encounter seasonal regimes, climate anomalies, market restructuring, and measurement revisions. Public agencies change eligibility rules, service channels, and administrative priorities. These changes can alter the marginal distribution of observed variables, the prevalence of the target, or the conditional relationship between evidence and outcome. A model can therefore remain available and computationally stable while the

decision process it represents becomes less reliable (Quiñonero-Candela *et al*., 2009; Gama *et al*., 2014; Webb *et al*., 2016; Lu *et al*., 2019).

Most operational controls focus on predictive deterioration. Accuracy, precision-recall performance, calibration, drift statistics, and threshold-specific errors are monitored, and a material change may trigger recalibration or retraining. Explainability is commonly treated as an attached artifact. A SHAP summary, permutation ranking, local surrogate, or counterfactual is validated during development and then reused as though its meaning remains stable. That assumption is unsafe. An explanation may be faithful to an old model but irrelevant to a changed environment, or faithful to an updated model while inconsistent with the reason previously communicated to decision makers. It may remain visually plausible while feature direction, subgroup reliability, and feasible recourse have changed (Ribeiro *et al*., 2016; Lundberg & Lee, 2017; Alvarez-Melis & Jaakkola, 2018; Rudin, 2019).

The institutional consequence is larger than a technical discrepancy. Explanations influence analyst attention, clinical review, customer communication, audit evidence, and appeal. A stable prediction with an unstable explanation can redirect human review toward the wrong evidence. A calibrated probability with an unfaithful explanation can make an adverse decision appear justified. The risk is particularly acute where geography, health status, socioeconomic context, or transaction behavior operate as proxies for vulnerable groups. Explanation availability does not itself establish causality, fairness, safety, or accountability (Barocas & Selbst, 2016; Selbst *et al*., 2019; Suresh & Guttag, 2021).

This study develops and evaluates the Adaptive Explainability Under Shift Framework, abbreviated AEUSF. The framework treats explanation reliability as a monitored lifecycle property. It links baseline validation, continuous shift detection, predictive and explanation-degradation assessment, uncertainty and subgroup review, intervention thresholds, joint adaptation, human escalation, approval, rollback, and organizational learning. The empirical contribution is a cross-sector temporal benchmark using four public data resources. Three settings contain naturally ordered regime changes. The healthcare setting uses real utilization covariates with a controlled shift because the available benchmark lacks event timestamps. Controlled and natural shifts are reported separately rather than blended into one claim.

The analysis answers ten research questions through a common design. It evaluates static and adaptive predictive strategies, four shift detectors, explanation fidelity and stability, the Explanation Degradation Index, uncertainty-instability association, subgroup explanation disparity, and an uncertainty-aware deferral policy. The results are intentionally bounded. The domain sample sizes and outcome semantics differ; the finance window is small; the energy series is an energy-climate planning signal rather than a grid-operational dataset; and the human-review analysis measures deferral workload rather than reviewer accuracy. These boundaries are part of the contribution because cross-sector governance should preserve common controls without pretending that one threshold is optimal everywhere.

2. Literature Review

Dataset-shift research distinguishes changes in the distribution that generates observed inputs and outcomes. Covariate-shift methods assume that the input distribution changes while the conditional target mechanism remains stable, whereas label-shift methods assume that class prevalence changes under a stable class-conditional input distribution (Sugiyama *et al*., 2007; Lipton *et al*., 2018). Concept drift concerns change in the posterior relationship used for prediction and can be sudden, gradual, incremental, or recurring. Reviews emphasize that detection, understanding, and adaptation are separate tasks and that a detector signal does not identify the cause of change (Žliobaitė, 2010; Moreno-Torres *et al*., 2012; Gama *et al*., 2014; Webb *et al*., 2016).

Distribution-shift evaluation has moved from univariate tests toward kernel, classifier-based, and embedding-based procedures. Maximum mean discrepancy tests whether two samples differ in a reproducing-kernel Hilbert space, while classifier two-sample tests estimate how easily a model distinguishes development data from current data (Gretton *et al*., 2012; Rabanser *et al*., 2019). In streaming settings, Page-Hinkley and ADWIN-style procedures detect sequential changes, but sensitivity introduces false alarms, delayed decisions, and retraining cost (Page, 1954; Bifet & Gavaldà, 2007). Benchmark work such as WILDS demonstrates that performance under real distribution shift cannot be inferred from random splits (Koh *et al*., 2021).

Explainable AI research offers complementary methods with different assumptions. LIME approximates a model around a focal point; SHAP allocates a prediction relative to a background distribution; permutation and ablation methods measure performance or score change after feature disruption; accumulated local effects summarize average response while reducing some extrapolation under dependence (Ribeiro *et al*., 2016; Lundberg & Lee, 2017; Apley & Zhu, 2020). Explanation evaluation has increasingly emphasized fidelity, robustness, stability, sanity checks, and the possibility that an explanation can be manipulated or insensitive to model changes (Adebayo *et al*., 2018; Alvarez-Melis & Jaakkola, 2018; Yeh *et al*., 2019; Slack *et al*., 2020).

The unresolved problem is temporal. Static explanation validation does not show that an explanation remains reliable when the reference distribution, model, or target mechanism changes. Counterfactual recourse may become infeasible when institutional rules or population constraints move. A local explanation can drift even when aggregate importance appears stable, and a stable explanation can remain attached to a poorly calibrated model. Uncertainty under distribution shift also behaves differently from explanation instability. Deep ensembles and conformal methods improve uncertainty characterization, but predictive uncertainty is not a direct confidence interval for an attribution (Gal & Ghahramani, 2016; Lakshminarayanan *et al*., 2017; Ovadia *et al*., 2019; Angelopoulos & Bates, 2023).

Sector-specific research reinforces the need for joint governance. Healthcare infrastructure work emphasizes security, privacy, continuity, and human authority (Hasan *et al*., 2022; Rasel *et al*., 2022; Kamruzzaman *et al*., 2023). Financial research links evolving fraud and lending systems to model risk, explainability, and fair-lending controls (Hasan

et al., 2023; Fahim *et al*., 2023; Rasel *et al*., 2023; Jahan *et al*., 2024). Energy and sustainability studies show that forecasting value depends on temporal validity, operational constraints, and environmental accounting rather than model complexity alone (Arman, Hasan, *et al*., 2024; Arman, Rasel, *et al*., 2024). Public-payment and climate-finance research further demonstrates that oversight, contestability, and distributional burden must be evaluated with technical performance (Rahat *et al*., 2024; Rabbani *et al*., 2024).

Calibration and class imbalance are central because high-risk decisions often use probabilities and thresholds rather than ranks alone. Brier score measures squared probability error, while post-hoc calibration can correct systematic overconfidence without changing rank order. Calibration curves and expected calibration error provide complementary diagnostics, and precision-recall analysis is preferred when positive events are uncommon (Brier, 1950; Niculescu-Mizil & Caruana, 2005; Guo *et al*., 2017; Saito & Rehmsmeier, 2015). The present study therefore reports discrimination, calibration, and threshold-specific outcomes separately.

Fairness under shift is equally multidimensional. Equal opportunity, predictive parity, and calibration express different objectives and can conflict when base rates differ. A temporal system may satisfy one criterion during development and violate it after population or policy change. Subgroup explanation disparity adds another layer because the model may provide less stable or less actionable reasons to one group even when aggregate error gaps remain small (Hardt *et al*., 2016; Chouldechova, 2017; Kleinberg *et al*., 2017; Mehrabi *et al*., 2021).

3. Theoretical Foundations

The theoretical model integrates distribution-shift theory with socio-technical governance. Distribution-shift theory explains why a previously validated mapping can lose relevance. Information-processing theory explains why organizations require monitoring capacity proportional to uncertainty and interdependence. Dynamic-capabilities theory adds the sequence of sensing, seizing, and reconfiguring. Organizational-learning theory explains how incident evidence and reviewer judgment become revised routines rather than isolated responses (Galbraith, 1974; March, 1991; Teece *et al*., 1997).

Explainability operates as an information-processing interface. It translates statistical behavior into evidence used by domain experts, risk officers, and affected parties. Its value depends on fidelity to the operative model, temporal stability, appropriate sensitivity, and suitability for action. Human-in-the-loop, human-on-the-loop, and human-in-command arrangements specify different allocations of decision rights. Automation research shows that oversight quality depends on authority, workload, feedback, and calibrated trust rather than the mere presence of a person (Parasuraman *et al*., 2000; Lee & See, 2004; Jarrahi, 2018; Raisch & Krakowski, 2021).

Institutional theory and algorithmic-accountability research place technical evidence inside a forum of answerability. An organization must be able to reconstruct the data, model version, explanation configuration, threshold, reviewer action, and remedy. Model cards, datasheets, and internal audits support this function, but documentation without authority or correction is incomplete (Bovens, 2007; Diakopoulos, 2016; Mitchell *et al*., 2019; Raji *et al*., 2020; Gebru *et al*., 2021). The AEUSF therefore treats monitoring and adaptation as necessary but not sufficient. Governance approval, rollback, contestability, and organizational learning are independent controls.

4. Distribution Shift and Explainability

Let X denote observed predictors and Y the target. Covariate shift occurs when P(X) changes while P(Y|X) is assumed stable. Prior-probability or label shift occurs when P(Y) changes while P(X|Y) is stable. Conditional distribution shift occurs when P(Y|X) changes and is the form most likely to invalidate a fixed decision rule. Domain shift is broader and often combines population, measurement, and conditional changes. Temporal drift identifies ordering, not mechanism. Concept drift is used here only when the predictive relationship changes, not as a synonym for any distributional difference.

The study evaluates controlled covariate and conditional shift in healthcare; natural prior and conditional shift in finance; natural temporal and conditional change in energy-climate conditions; and natural prior and population change in the public-sector panel. Abrupt and gradual changes are represented by transitions between blocked periods. Recurrence is not formally identified in the short windows. This classification prevents a detector statistic from being interpreted as proof of concept change.

Explainability can degrade through several pathways. A changed reference distribution alters the baseline against which feature contributions are interpreted. Retraining changes functional dependence and feature rank. Calibration changes probabilities without changing feature effects, so it may improve Brier score while leaving explanation quality unchanged. Population shift can also change which subgroups receive unstable explanations. The practical unit is therefore not a single importance plot but a temporal evidence bundle containing prediction, calibration, fidelity, ranking, direction, actionability, subgroup disparity, and uncertainty.

Table 1: Definitions and taxonomy of distribution shift

Shift type	Mathematical distinction	Operational interpretation
Covariate shift	P(X) changes; P(Y|X) assumed stable	Population, channel, or measurement change
Label shift	P(Y) changes; P(X|Y) assumed stable	Changing prevalence or policy threshold
Conditional shift	P(Y|X) changes	Mechanism, behavior, or institutional process changes
Concept drift	Time-indexed conditional change	Sudden, gradual, incremental, or recurring
Domain shift	Source and target environments differ	Combined population and measurement differences
Policy-induced shift	Rules alter observed behavior or labels	Eligibility, investigation, coding, or treatment change

Note: P denotes a probability distribution, X predictors, and Y the target.

5. Adaptive Explainability Under Shift Framework

The AEUSF contains fifteen lifecycle stages. Baseline model validation and baseline explanation validation establish approved predictive and explanatory configurations. Continuous monitoring then evaluates data, model outcomes, explanations, uncertainty, and subgroup behavior. A shift signal triggers separate predictive-performance and explanation-degradation assessments. Uncertainty and fairness reviews determine whether the observed change is broadly distributed or concentrated in high-risk cases. Intervention thresholds map evidence to a response rather than treating every detector alert as a retraining instruction.

Adaptation separates model recalibration or retraining from explanation recalculation. Recalibration may be sufficient when score meaning changes while ranking and feature dependence remain stable. Retraining is appropriate when relationships change materially. Explanation recalculation updates reference data, attribution summaries, local evidence, and actionability checks. Human review examines cases and affected groups when EDI, uncertainty, calibration, or subgroup thresholds fail. Governance approval records the evidence, responsible owner, approved duration, review deadline, and rollback condition. Post-intervention monitoring tests whether the response solved the problem or merely moved it.

The framework is deliberately sector dependent. A clinical false negative, a financial false positive, an energy forecast error, and a public-benefit eligibility error do not have the same consequence or review deadline. AEUSF provides a common control architecture while allowing risk limits, explanation detail, acceptable deferral, and human authority to differ. This design aligns with the NIST AI Risk Management Framework, ISO risk-management and management-system standards, model-risk guidance, health-AI guidance, and the risk-tiered obligations of the European Union AI Act, all available before the evidence cutoff (Board of Governors of the Federal Reserve System & Office of the Comptroller of the Currency, 2011; World Health Organization, 2021; NIST, 2023; ISO, 2023a, 2023b; European Union, 2024).

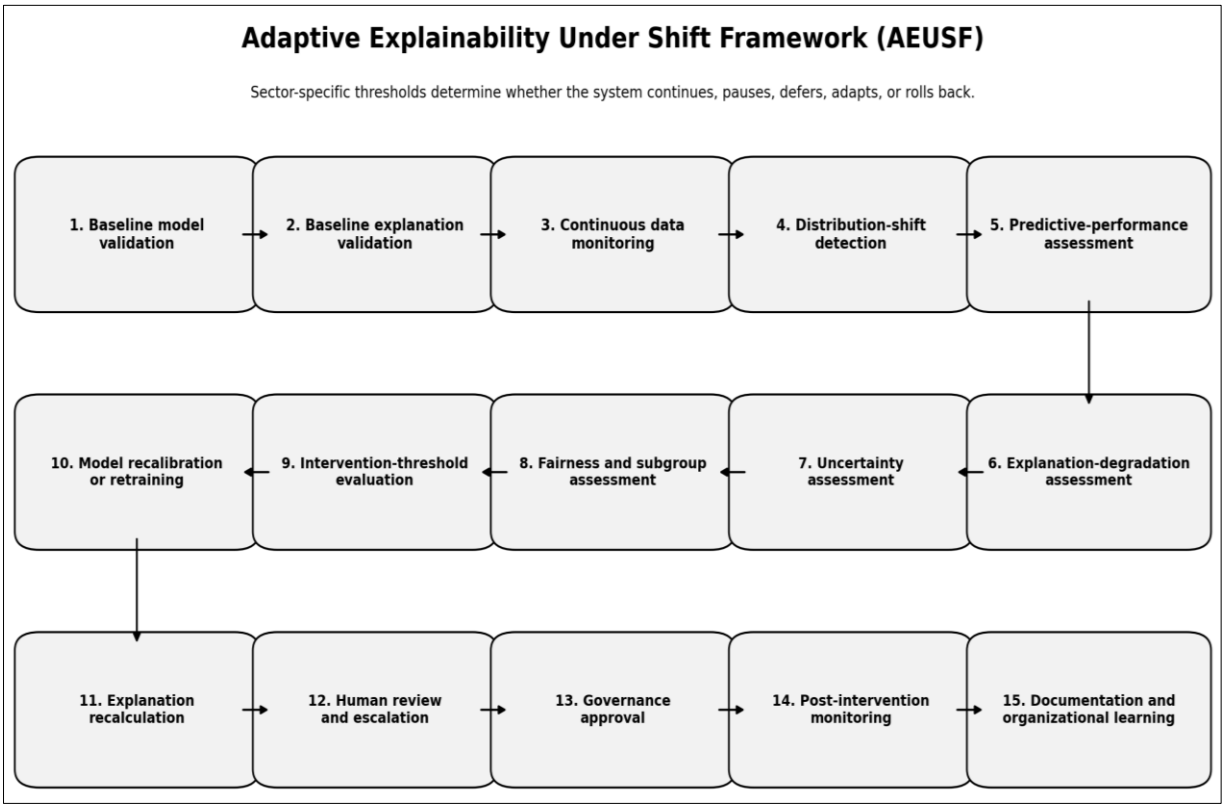

Fig 1: Adaptive Explainability Under Shift Framework with fifteen lifecycle stages.

6. Research Questions and Hypotheses

The study asks how predictive performance, explanation reliability, uncertainty, subgroup disparity, and operational workload change across temporal environments. RQ1 concerns predictive degradation. RQ2 to RQ5 examine fidelity, stability, recalibration, and adaptive explanation updating. RQ6 and RQ7 compare sectors and subgroups. RQ8 tests whether predictive uncertainty identifies explanation instability. RQ9 translates evidence into intervention thresholds. RQ10 specifies the governance mechanisms required after deployment.

Six hypotheses are empirically testable with the available design. H1 predicts lower predictive performance during shift. H2 predicts lower explanation fidelity. H3 predicts lower temporal rank stability as shift magnitude rises. H4 predicts that recalibration improves calibration but does not restore explanation reliability. H5 predicts that joint model and explanation adaptation improves decision reliability more than static operation. H6 predicts a positive relationship between predictive uncertainty and explanation instability. Claims concerning actual human decision improvement and causal sector moderation are treated as propositions for future field research because no participant study or sector-randomized design is available.

7. Research Methodology

The study uses a comparative temporal benchmark. Each domain is divided into training, validation, pre-shift, shift,

and post-shift windows. All preprocessing parameters are fitted on training observations. Validation selects classification thresholds. Future observations are never used to transform earlier observations. The same analytical sequence is applied across domains: data audit, static modeling, shift measurement, explanation analysis, EDI calculation, subgroup analysis, uncertainty analysis, adaptive strategy comparison, and governance interpretation.

Three predictive models provide different complexity levels: standardized logistic regression, random forest, and gradient boosting. Random forest is the primary explanation model because it captures nonlinear dependence and supports direct score-ablation analysis without assuming an attribution background distribution. The adaptive comparison uses logistic regression to preserve architecture-matched retraining and recalibration across limited windows. This choice avoids attributing a model-family advantage to adaptation policy.

Statistical analysis emphasizes repeated domain-period contrasts rather than pooled observation-level significance. Bootstrap intervals quantify uncertainty for random-forest PR-AUC, ROC-AUC, and Brier score. Cross-domain pre-to-shift changes use paired Wilcoxon tests with standardized mean differences. Adaptive strategies are compared across eight shift and post-shift tasks with a Friedman test and Holm-adjusted pairwise Wilcoxon tests. The number of domains is small, so effect sizes and pattern consistency receive more weight than conventional significance labels.

8. Data Sources and Preprocessing

The healthcare analysis uses the RAND Health Insurance Experiment utilization benchmark. A deterministic sample of 4,000 records was selected from real healthcare covariates. Because the public benchmark lacks event time, the analysis does not claim a naturally observed healthcare drift. It applies a documented controlled covariate and conditional shift to later windows while preserving the observed feature distribution as the empirical base. The target is a controlled high-utilization event, and the subgroup indicator represents poor reported health. This design tests mechanism sensitivity rather than clinical deployment.

The finance analysis uses quarterly U.S. macroeconomic series ordered through time. High inflation, defined before model comparison as inflation above 4.5 percent, is the target. The setting represents policy and financial-risk monitoring rather than an individual lending decision. Regime transitions generate natural changes in prevalence and predictor relationships. The sample is only 202 quarters, and the evaluation windows contain few positive events. Results are therefore diagnostic and accompanied by wide bootstrap intervals.

The energy-climate analysis uses 720 monthly El Nino sea-surface-temperature observations as a planning signal relevant to renewable-energy and demand uncertainty. The target is a warm anomaly above 0.75 standardized units. Seasonal and lagged features preserve time order. This is not a grid-failure dataset, and the manuscript does not infer operational reliability directly. The public-sector analysis uses 1,562 Gapminder country-year observations. The target, life expectancy below 60 years, represents a resource-prioritization vulnerability signal rather than an administrative eligibility decision.

Across domains, missing values are imputed with training medians, continuous predictors are standardized only where required, and categorical or seasonal fields are represented through documented indicators. Subgroups are defined from observed and ethically interpretable fields: poor-health status, high-unemployment regime, peak-season months, and lower-income country-years. They are analytical monitoring groups, not claims about protected identity.

The data provenance is documented rather than inferred from software packaging. The healthcare benchmark derives from the RAND Health Insurance Experiment, whose design and limitations are described by Newhouse and the Insurance Experiment Group (1993). The sea-surface-temperature series is interpreted using the established El Nino definition and seasonal context described by Trenberth (1997). The country-year panel is drawn from Gapminder's documented development indicators (Gapminder Foundation, 2024). These sources support transparent reuse, but the analytical targets created for this benchmark remain the responsibility of the present study.

Table 2: Dataset characteristics and temporal partitions

Domain	Source	N	p	Train N	Validation N	Pre N	Shift N	Post N	Event rate	Shift form	Subgroup
Healthcare	RAND Health Insurance Experiment utilization data	4000	9	2000	400	400	600	600	0.201	Controlled covariate and conditional shift on real healthcare covariates	Poor-health indicator
Finance	US macroeconomic quarterly data	202	11	83	20	20	40	39	0.322	Natural prior and conditional shift across inflation regimes	High-unemployment regime
Energy and climate	El Nino sea-surface temperature series used as an energy-climate planning signal	720	8	348	60	60	120	132	0.285	Natural temporal and conditional shift in ocean temperature anomalies	Peak-season months
Public sector	Gap minder country-year development indicators	1562	8	710	142	142	284	284	0.461	Natural prior and population shift in development vulnerability	Lower-income country-years

Note: Healthcare uses controlled shift on real covariates. The other domains use natural temporal ordering. Event definitions are domain-specific and are not directly comparable.

9. Distribution-Shift Identification

Shift identification combines domain knowledge and statistical evidence. Population Stability Index and maximum univariate Kolmogorov-Smirnov distance detect marginal changes. Mean Wasserstein distance summarizes scaled univariate movement. Maximum mean discrepancy captures multivariate difference, and a domain classifier estimates whether current observations can be separated from training observations. Thresholds were declared before comparative interpretation: PSI at 0.20, maximum KS at 0.20, domain-

classifier AUC at 0.70, and MMD at 0.02.

These thresholds are operating points, not universal scientific constants. Their purpose is to create comparable alert behavior across domains. False alarms are windows flagged despite being designated pre-shift; missed shifts are designated shift windows not flagged. Sensitivity analysis varies the thresholds and shows the expected trade-off between detection and alert burden. A detector alarm opens diagnosis. It does not establish that P(Y|X) changed, nor does it authorize retraining by itself.

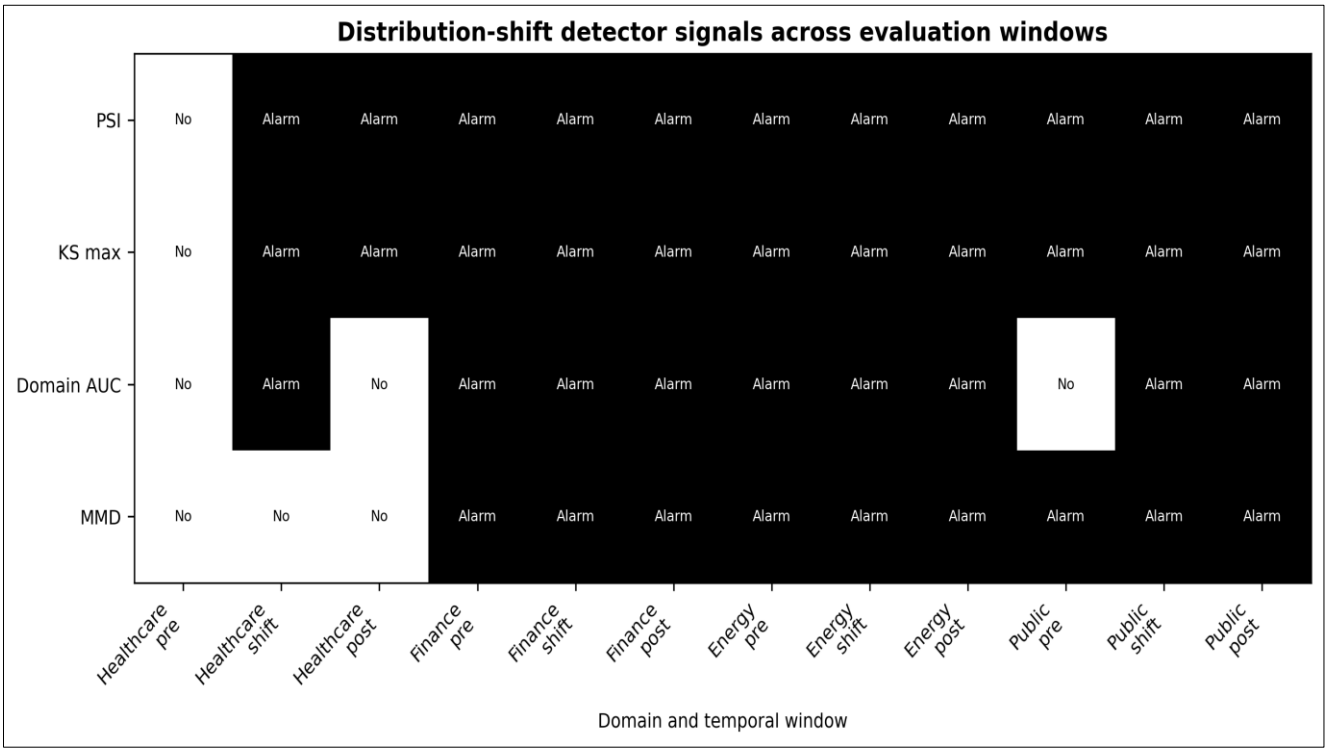

Fig 2: Distribution-shift detection timeline across domains and periods.

10. Predictive and Adaptive Models

Static models are trained once on the development window. Recalibration updates probability mapping without changing model coefficients or feature relationships. Periodic and expanding strategies add recent labeled observations at fixed evaluation points. Rolling retraining discards older observations and therefore trades stability for responsiveness. Drift-triggered retraining uses the declared detector rule. Joint adaptation retrains the logistic model, recomputes the explanation reference, and defers the 20 percent of cases with highest predictive entropy to human review.

The joint strategy is a governed selective-prediction policy rather than a claim that abstention improves every metric. Deferred cases are removed from automatic decisions and counted as review workload. The policy is evaluated on the remaining cases and reports review rate. In deployment, the benefit would depend on reviewer capacity and accuracy, which are not observed here. The analysis therefore interprets deferral as a workload and risk-routing mechanism, not evidence of improved human judgment (Geifman & El-Yaniv, 2017).

11. Explainability and Explanation-Degradation Measures

Feature-ablation explanations are calculated by replacing one feature at a time with its training median and recording the change in the random-forest score. Global importance is mean absolute score change. Deletion fidelity measures how much of the score distance from a low-risk reference is removed when the highest-ranked features are replaced. Temporal rank stability is normalized Spearman agreement between pre-shift and current rankings. Top-five Jaccard overlap is reported separately. Directional consistency is the share of commonly important features retaining contribution sign. Method agreement compares ablation rank with native forest importance.

Actionability is the share of total attribution assigned to fields classified as mutable or institutionally controllable. It is not recourse. A mutable field may remain costly, ethically inappropriate, or outside the affected person's control. Subgroup explanation disparity is the normalized difference in mean attribution magnitude between the monitored groups. Uncertainty is normalized predictive entropy. Sparsity is the number of features needed to account for 80 percent of attribution.

The Explanation Degradation Index ranges from zero to one. It is the equal-weight mean of six normalized loss components: one minus fidelity, one minus Spearman rank stability, one minus directional consistency, subgroup disparity, predictive uncertainty, and one minus actionability. Equal weighting is used because no empirical basis justifies allowing one dimension to dominate. Fidelity-heavy, fairness-heavy, and uncertainty-heavy schemes test sensitivity. EDI is a screening index. It does not certify that an explanation is legally sufficient or causally correct.

Counterfactual and recourse methods are not used as the common quantitative basis because their validity depends on domain-specific feasibility, causal structure, and actor control. They remain important governance checks. A mathematically close counterfactual can be impossible, impermissible, or misleading when dependencies change under shift. AEUSF therefore requires action-set review and

distinguishes a model-valid contrast from a real-world intervention (Wachter *et al*., 2017; Poyiadzi *et al*., 2020; Karimi *et al*., 2021). Broader surveys similarly caution that explanation methods answer different questions and should not be evaluated through one fidelity measure alone (Guidotti *et al*., 2018).

12. Empirical Results

The four settings differ sharply in sample size, event prevalence, and predictive difficulty. Healthcare contains 4,000 observations and an event rate of 0.201. Finance contains 202 quarters and an event rate of 0.322, but its temporal windows include only two shift-period and four post-period events. Energy-climate contains 720 monthly observations with an event rate of 0.285. Public-sector data contain 1,562 country-years with an event rate of 0.461. These differences justify within-domain temporal comparisons rather than pooled accuracy claims.

Static random-forest performance does not show a single shift pattern. Healthcare PR-AUC rises from 0.257 before shift to 0.384 during shift and 0.398 afterward because the controlled outcome mechanism increases separability. Finance PR-AUC falls from 0.497 to 0.243 and then 0.156, while the shift-period ROC-AUC rises to 0.882 because two events happen to be ranked highly. Energy-climate PR-AUC declines from 0.975 to 0.926 and 0.758. Public-sector PR-AUC remains high at 0.991, 0.979, and 0.987. Bootstrap intervals show that the finance estimates are highly uncertain. Calibration also moves independently from ranking. Random-forest Brier score changes from 0.221 to 0.224 in healthcare, from 0.345 to 0.273 in finance, from 0.062 to 0.088 in energy-climate, and from 0.034 to 0.039 in the public sector. The cross-domain Wilcoxon test for shift-minus-pre PR-AUC yields a mean difference of -0.047, p=0.625, standardized effect -0.298. With only four domains, this is evidence of heterogeneous direction rather than a null claim about all high-risk systems.

Detector sensitivity is high but specificity is poor at the declared thresholds. PSI and maximum KS detect all eight designated shift/post windows but each generates three pre-window false alarms. Domain-classifier AUC detects seven of eight with two false alarms. MMD detects six of eight with three false alarms. The result supports a multi-signal confirmation rule rather than automatic retraining.

Explanation reliability changes differently from predictive performance. Healthcare EDI changes from 0.355 to 0.361 during shift and 0.345 afterward. Finance EDI is 0.344, 0.342, and 0.386. Energy-climate EDI rises from 0.242 to 0.294 and 0.312. Public-sector EDI rises from 0.262 to 0.334 before declining to 0.304. Mean shift-period EDI increases by 0.032 across domains, p=0.250, with a standardized effect of 0.892. Rank stability declines by 0.073 on average, p=0.125, standardized effect -0.744. These large standardized changes are imprecise but operationally relevant.

Recalibration frequently improves Brier score without changing EDI. The largest improvements occur in healthcare shift (-0.083) and post (-0.072), finance shift (-0.066), and finance post (-0.091). Because recalibration leaves the model's feature dependence and explanation reference unchanged in this design, EDI remains unchanged by construction. This directly supports the distinction between probability repair and explanation repair.

Across eight domain-period tasks, the Friedman test rejects equal PR-AUC distributions across strategies, statistic 19.024, p=0.004. Pairwise Holm-adjusted comparisons against static operation are not significant. Joint adaptation plus uncertainty deferral improves mean PR-AUC by 0.020 and reviews about 20 percent of cases. It improves energy-climate post-period PR-AUC by 0.089 and finance post-period PR-AUC by 0.073, but reduces finance shift-period PR-AUC by 0.100. Selective adaptation therefore creates useful options without supporting unconditional superiority.

Predictive uncertainty is not a general detector of explanation instability. Spearman correlations are 0.277 in healthcare, 0.039 in finance, 0.018 in energy-climate, and -0.800 in the public-sector sample. The negative public-sector result reflects a setting where highly confident predictions can still rely on explanations that vary across bootstrap models. H6 is not supported across domains. Explanation uncertainty requires direct measurement.

Table 3: Random-forest predictive performance before and after shift

Domain	Period	N	Event rate	PR-AUC	ROC-AUC	Brier	F1	MCC
Energy and climate	post	132	0.341	0.758	0.846	0.160	0.636	0.455
Energy and climate	pre	60	0.283	0.975	0.986	0.062	0.938	0.918
Energy and climate	shift	120	0.383	0.926	0.961	0.088	0.879	0.805
Finance	post	39	0.103	0.156	0.396	0.204	0.200	0.090
Finance	pre	20	0.350	0.497	0.648	0.345	0.560	0.245
Finance	shift	40	0.050	0.243	0.882	0.273	0.333	0.397
Healthcare	post	600	0.217	0.398	0.673	0.217	0.414	0.208
Healthcare	pre	400	0.182	0.257	0.620	0.221	0.310	0.109
Healthcare	shift	600	0.195	0.384	0.695	0.224	0.400	0.210
Public sector	post	284	0.313	0.987	0.992	0.032	0.947	0.927
Public sector	pre	142	0.352	0.991	0.994	0.034	0.950	0.923
Public sector	shift	284	0.331	0.979	0.985	0.039	0.909	0.873

Note: PR-AUC is the primary ranking metric; Brier assesses probability accuracy.

Table 4: Explanation Degradation Index results

Domain	Period	EDI	Uncertainty	Actionability
Healthcare	pre	0.355	0.957	0.606
Healthcare	shift	0.361	0.964	0.487
Healthcare	post	0.345	0.953	0.567
Finance	pre	0.344	0.872	0.684
Finance	shift	0.342	0.857	0.661
Finance	post	0.386	0.835	0.775
Energy and climate	pre	0.242	0.501	0.426
Energy and climate	shift	0.294	0.594	0.382
Energy and climate	post	0.312	0.598	0.417
Public sector	pre	0.262	0.270	0.156
Public sector	shift	0.334	0.287	0.133
Public sector	post	0.304	0.266	0.139

Note: EDI ranges from zero to one; higher values indicate greater degradation.

Table 5: Adaptive-strategy comparison across eight domain-period tasks

Strategy	PR-AUC	Brier	F1	EDI	Review_rate	Abstention_rate
Drift-triggered retrain	0.616	0.125	0.591	0.390	0.414	0.000
Expanding retrain	0.616	0.125	0.591	0.390	0.414	0.000
Joint adaptation plus deferral	0.644	0.106	0.623	0.390	0.201	0.201
Periodic retrain	0.616	0.125	0.591	0.390	0.414	0.000
Recalibration only	0.624	0.095	0.487	0.354	0.207	0.000
Rolling retrain	0.612	0.131	0.580	0.412	0.333	0.000
Static	0.624	0.134	0.583	0.354	0.372	0.000

Note: Metrics are unweighted task means. Review rate includes model-positive review for ordinary strategies and uncertainty deferral for the joint strategy.

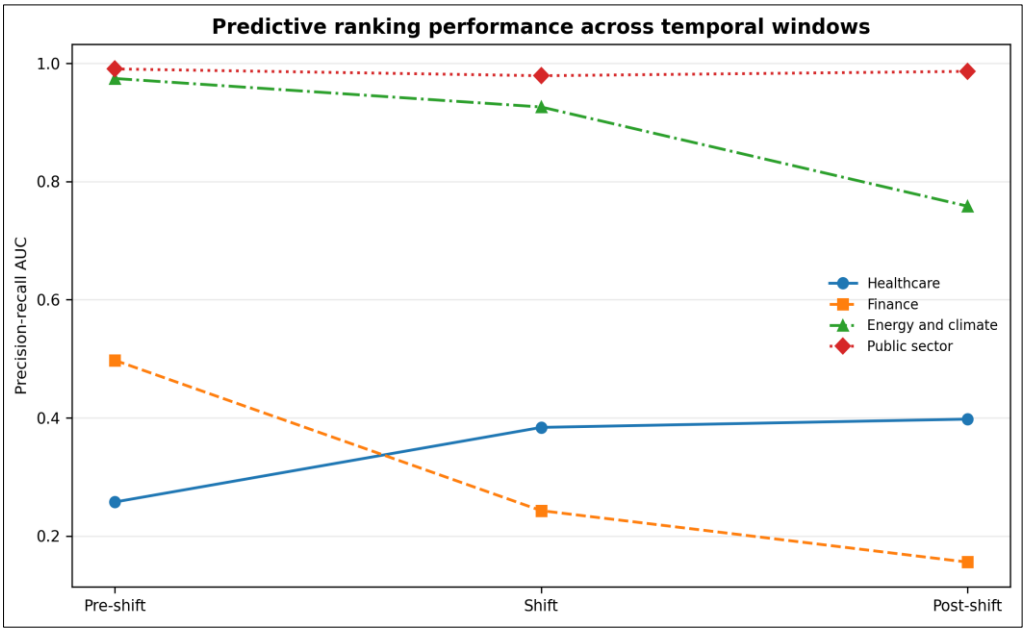

Fig 3: Predictive-performance degradation over time by domain.

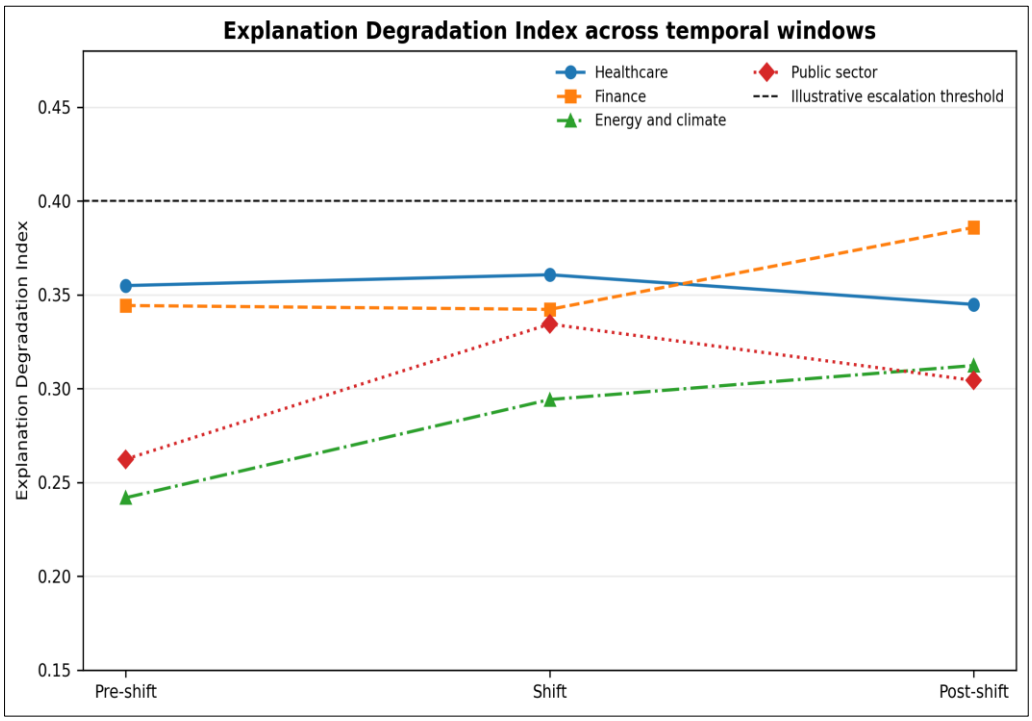

Fig 4: Explanation Degradation Index over time.

13. Cross-Sector Analysis

Cross-sector comparison reveals three governance patterns. First, detectability is not equivalent to harm. Finance and energy-climate pre-windows are statistically separable from training data, producing false alarms even before the designated shift period. This reflects long temporal distance and seasonality rather than necessarily harmful concept change. Second, explanation degradation is most visible where contextual relations change or where a high-performing model becomes less discriminative. Energy-climate and public-sector EDI rise during shift even though their absolute predictive performance remains stronger than healthcare and finance. Third, uncertainty and explanation instability do not align consistently.

Healthcare requires review of clinical meaning and patient burden. The controlled experiment shows that an apparently improved PR-AUC can coexist with explanation degradation, so improved discrimination cannot waive explanation review. Finance requires stricter temporal validation, calibration, and queue-capacity controls because few events can destabilize both metrics and explanations. Energy and climate planning requires attention to seasonal and regime-specific reference distributions. Public-sector systems require due-process review when a stable aggregate model creates unequal explanation burden across income or regional groups.

The common architecture is continuous monitoring, direct explanation-quality assessment, uncertainty and subgroup review, human escalation, versioned adaptation, and rollback. Sector rules change thresholds, deadlines, documentation detail, and the acceptable balance of false positives and false negatives. A shared technical platform can support these controls, but accountability remains with named institutional decision owners.

14. Discussion

The empirical evidence rejects a simple chain in which distribution shift necessarily reduces predictive performance, causes explanation degradation, and is corrected by retraining. The observed pathways are conditional. Healthcare becomes easier to discriminate under the controlled mechanism while explanation actionability and subgroup disparity change. Finance shows unstable rankings with very small event counts. Energy-climate shows persistent predictive value but rising EDI. Public-sector prediction remains strong while explanation disparity increases. These patterns show why explanation monitoring cannot be proxied by an AUC dashboard.

The EDI provides a practical synthesis but should not be treated as a universal score. Equal weighting preserves visibility across fidelity, stability, direction, disparity, uncertainty, and actionability. Weight sensitivity shows that absolute values move when fidelity, fairness, or uncertainty receives greater emphasis, while the comparative temporal pattern is more stable. Organizations should therefore govern the dimension profile and use the total as a triage signal, not as a certification label.

Recalibration offers the clearest conceptual result. It can improve probability quality without changing the explanatory relationship. This is desirable when threshold policy depends on probability meaning, but insufficient when reason-giving, recourse, or analyst workflow depends on the current feature structure. Model and explanation updates should be approved as distinct configuration items. The joint strategy also demonstrates the value and cost of deferral. It improves some tasks and creates a clear review channel, but it does not dominate, and its 20 percent workload may be infeasible in real-time systems.

Detector behavior reinforces the governance argument. Marginal detectors are sensitive but generate false alarms.

A multivariate domain classifier is more selective but still confuses benign temporal difference with harmful shift. The appropriate response is a confirmation stage that combines distributional evidence, outcome evidence, explanation evidence, data-quality checks, policy context, and domain intelligence. Automated monitoring should accelerate diagnosis, not replace it.

15. Theoretical Contributions

The study contributes a temporal theory of explainability. Explainability is not a fixed attribute of a model artifact; it is a relationship among the current model, current population, explanation reference, decision context, and institutional use. Distribution shift can affect any part of that relationship. This view extends XAI beyond static fidelity and extends concept-drift research beyond predictive adaptation.

The AEUSF also connects dynamic capabilities with algorithmic accountability. Sensing includes data, model, explanation, uncertainty, and fairness signals. Seizing involves selecting recalibration, retraining, explanation update, deferral, or no action. Reconfiguration requires validation, approval, deployment control, rollback, and learning. Accountability determines who may authorize each step and what evidence must be retained. The resulting theory explains why two organizations using the same detector and model can create different risk outcomes.

Finally, the study separates prediction confidence from explanation confidence. The cross-domain uncertainty results show that entropy cannot substitute for direct explanation-instability measurement. Selective prediction remains valuable, but an explanation-specific risk measure is needed to route cases in which a confident model offers an unstable reason.

16. Practical and Managerial Implications

Managers should begin with an inventory of consequential systems and an approved explanation baseline. The baseline should record data period, model and threshold version, reference distribution, global and local explanation method, actionability rules, subgroup metrics, and known limitations. Monitoring should then compare current evidence against that baseline at a frequency proportionate to decision speed and label delay.

Threshold design should be multi-signal. A high shift statistic with stable outcomes and explanations may justify investigation without retraining. Calibration failure with stable rank and explanations may justify recalibration. EDI failure should trigger explanation recomputation and case review even when discrimination is stable. A severe subgroup disparity or infeasible counterfactual should escalate independently of the aggregate EDI. Managers should also estimate reviewer workload before approving deferral.

The implementation roadmap begins with inventory and decision rights, moves to monitoring and threshold testing, then validates adaptive controls and develops organizational learning. A small initial scope is preferable. Expanding before reviewers, rollback assets, and documentation are stable converts technical responsiveness into operational fragility.

Human-AI interaction design should make uncertainty, evidence provenance, and the ability to disagree visible at the point of use. Guidelines for human-AI interaction emphasize expectation setting, efficient correction, and support for user control (Amershi *et al*., 2019). Operational leaders should also budget for hidden technical debt: monitoring pipelines, feature definitions, explanation references, and retraining dependencies create maintenance obligations that are easily omitted from model-performance estimates (Sculley *et al*., 2015).

17. Governance and Policy Implications

Policy should distinguish the validation of a model from the continuing validity of its explanations. High-risk systems should document explanation reference data, monitoring metrics, materiality thresholds, review authority, update approval, and rollback. A requirement to provide explanations without a requirement to monitor their degradation can institutionalize misleading reasons. Standards and supervisory frameworks should therefore treat explanation configuration as a controlled model component.

Human oversight must be defined through authority and evidence. Human in the loop is appropriate when every consequential decision requires approval. Human on the loop is appropriate when automation operates within bounded limits and a supervisor can intervene. Human in command requires senior authority to set autonomy limits, suspend operation, and determine acceptable residual risk. The appropriate arrangement depends on severity, reversibility, time pressure, and the reliability of evidence.

Audit logs should connect data snapshot, detector signal, predictive metrics, explanation metrics, EDI components, subgroup results, adaptation decision, reviewer identity, approval, deployment time, and post-change outcome. Public-sector and consumer-facing systems also need contestability. An affected party should be able to challenge factual inputs and receive reconsideration rather than a static explanation generated from a superseded model.

The policy architecture is consistent with authoritative pre-cutoff guidance. The National Institute of Standards and Technology (2023) frames AI risk as a lifecycle governance responsibility. The International Organization for Standardization (2023a, 2023b) separates risk-management guidance from management-system requirements. The World Health Organization (2021) emphasizes human autonomy, accountability, and public-interest safeguards in health AI, while the European Union (2024) establishes risk-tiered obligations for high-risk systems. These instruments do not provide a universal EDI threshold, so sector validation remains necessary.

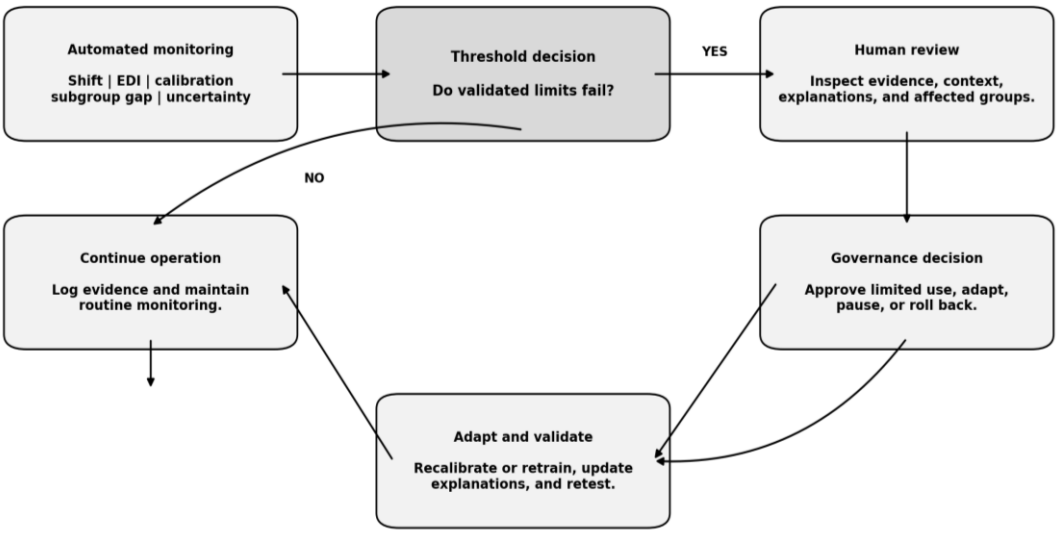

Fig 5: Human-escalation workflow for unreliable explanations.

18. Limitations

The first limitation is representativeness. The four public resources are benchmarks rather than production decision systems. Healthcare uses a controlled shift on real covariates because timestamps are unavailable. Finance contains few positive events in later windows. The energy-climate series is a planning signal rather than a grid operation dataset. The public-sector target is a vulnerability indicator, not an actual benefit or inspection decision. These choices permit transparent cross-sector comparison but restrict external validity.

Second, explanation quality is operationalized through feature ablation and related ranking metrics. Results may differ for conditional SHAP, local surrogates, counterfactual recourse, integrated gradients, or graph explanations. Actionability is a feature classification rather than evidence that a person can feasibly change the factor. Subgroup disparity uses available monitoring groups and does not establish discrimination or legal protected status.

Third, the study does not include human participants. Deferral and review rates measure operational workload, not reviewer accuracy, trust calibration, comprehension, or procedural justice. Human escalation is a governance design inferred from technical evidence and prior research. Field and laboratory validation are required before claiming improved human-AI decisions.

Fourth, EDI thresholds and equal weights are normative operating choices. Sensitivity analysis reduces but does not remove that dependence. The index can conceal offsetting dimensions, so component scores must accompany the total. Detector thresholds are also sample dependent. A larger monitoring window would reduce noise but delay action.

Finally, the numerical benchmark was executed in a contemporary container using algorithms and APIs that existed before the evidence cutoff. The supplementary package provides a pre-cutoff compatibility specification, but exact bitwise equivalence across software versions is not claimed. This disclosure preserves the historical evidence boundary without pretending that the execution environment itself was frozen in May 2025.

19. Future Research

Future research should evaluate AEUSF on longitudinal clinical, payment, energy, and public-administration systems with actual model updates, appeal outcomes, reviewer actions, and incident records. Strong designs would compare static explanations, periodically updated explanations, drift-triggered updates, and uncertainty-aware deferral under randomized or stepped-wedge deployment. Human outcomes should include error correction, calibrated override, decision time, written justification quality, and burden across affected groups.

Methodological work should develop explanation-specific uncertainty intervals and conformal procedures for attribution sets. Causal and counterfactual explanations should be tested under changing structural assumptions, not only changing model parameters. Dynamic reference selection for SHAP and local surrogates requires governance because updating the reference can change reasons even when the model is fixed.

Cross-sector research should estimate how decision severity, reversibility, and time pressure moderate thresholds. Clinical triage may require rapid human review with high sensitivity. Fraud monitoring may tolerate temporary deferral but face strategic gaming. Energy operations may require low-latency fallback rules. Public administration may require stronger reason-giving and contestability even when response time is slower. A mature theory should link these institutional characteristics to validated monitoring and intervention limits.

20. Conclusion

Distribution shift changes more than predictive performance. It can alter which features appear important, how stable explanations remain, whether directions persist, which groups receive unreliable reasons, and whether a proposed action is meaningful. The cross-sector results show that these changes do not move together. Predictive performance can improve while explanation quality worsens, and recalibration can repair probabilities while leaving explanation degradation untouched.

The AEUSF responds by making explanation reliability a lifecycle control. It separates detection from diagnosis, model adaptation from explanation recalculation, uncertainty from explanation confidence, and technical update from governance approval. The EDI provides a transparent screening measure whose components remain visible. Joint adaptation and deferral can improve some tasks, but the mixed results rule out automatic retraining or universal thresholds.

The practical standard is controlled adaptation. High-risk organizations should monitor data, predictions, explanations, uncertainty, and subgroup outcomes; escalate when validated limits fail; preserve human authority; and document every material change. Explainability is useful only while the institution can show that the explanation still describes the operative decision process and remains suitable for the people who must rely on, challenge, or correct it.

References

1. Adebayo J, Gilmer J, Muelly M, Goodfellow I, Hardt M, Kim B. Sanity checks for saliency maps. Adv Neural Inf Process Syst. 2018;31.
2. Alvarez-Melis D, Jaakkola TS. On the robustness of interpretability methods. arXiv. 2018. doi:10.48550/arXiv.1806.08049.
3. Amershi S, Weld D, Vorvoreanu M, Fourney A, Nushi B, Collisson P, *et al*. Guidelines for human-AI interaction. Proc 2019 CHI Conf Hum Factors Comput Syst. 2019:1-13. doi:10.1145/3290605.3300233.
4. Angelopoulos AN, Bates S. Conformal prediction: A gentle introduction. Found Trends Mach Learn. 2023;16(4):494-591. doi:10.1561/2200000101.
5. Apley DW, Zhu J. Visualizing the effects of predictor variables in black box supervised learning models. J R Stat Soc Series B Stat Methodol. 2020;82(4):1059-86. doi:10.1111/rssb.12377.
6. Arman M, Hasan MN, Rasel IH. Clean energy transition in USA: Big data analytics for renewable energy forecasting and carbon reduction. J Manag World. 2024;(3):192-206. doi:10.53935/jomw.v2024i4.1196.
7. Arman M, Rasel IH, Razib MNH, Fahim ASM. Big data and machine learning for sustainable waste reduction. J Posthumanism. 2024;4(2):448-67. doi:10.63332/joph.v4i2.3361.
8. Barocas S, Selbst AD. Big data's disparate impact. Calif Law Rev. 2016;104(3):671-732.
9. Bifet A, Gavaldà R. Learning from time-changing data with adaptive windowing. In: Proc SIAM Int Conf Data Min. 2007. p. 443-8. doi:10.1137/1.9781611972771.42.
10. Board of Governors of the Federal Reserve System, Office of the Comptroller of the Currency. Supervisory guidance on model risk management (SR 11-7/OCC Bulletin 2011-12). 2011.
11. Bovens M. Analysing and assessing accountability: A conceptual framework. Eur Law J. 2007;13(4):447-68. doi:10.1111/j.1468-0386.2007.00378.x.
12. Brier GW. Verification of forecasts expressed in terms of probability. Mon Weather Rev. 1950;78(1):1-3.
13. Chouldechova A. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. Big Data. 2017;5(2):153-63. doi:10.1089/big.2016.0047.
14. Diakopoulos N. Accountability in algorithmic decision making. Commun ACM. 2016;59(2):56-62. doi:10.1145/2844110.
15. European Union. Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence. Off J Eur Union. 2024.
16. Fahim ASM, Ibrahim M, Pritty AA, Tania TA. Algorithmic accountability in U.S. consumer FinTech: Governance mechanisms for credit risk, fair lending, and financial stability. J Econ Finance Account Stud. 2023;5(4):80-93. doi:10.32996/jefas.2023.5.4.8.
17. Gal Y, Ghahramani Z. Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In: Proc 33rd Int Conf Mach Learn. 2016. p. 1050-9.
18. Galbraith JR. Organization design: An information processing view. Interfaces. 1974;4(3):28-36. doi:10.1287/inte.4.3.28.
19. Gama J, Žliobaitė I, Bifet A, Pechenizkiy M, Bouchachia A. A survey on concept drift adaptation. ACM Comput Surv. 2014;46(4):Article 44. doi:10.1145/2523813.
20. Gapminder Foundation. Gapminder data documentation and country-year indicators. Gapminder Foundation; 2024.

How to Cite This Article

Reynolds PM, Cruz VB. Explainability under distribution shift: adaptive explainable artificial intelligence for high-risk decision systems. Int J Artif Intell Eng Transform. 2025;6(1):13-24. doi:10.54660/IJAIET.2025.6.1.13-24.

Creative Commons (CC) License