

DOI: https://doi.org/10.48009/3_iis_2026_101

Balancing accuracy and transparency: When explainable AI enhances decision-making in high-stakes environments

Taylor Stonelake, *Marymount University*, and tcm05307@marymount.edu

Abstract

It's becoming clear that AI is going to play a major role in the decisions formerly made by individuals. Although accuracy of AI systems' decisions is an essential aspect of AI systems' functioning, when an AI system makes a decision that has real-world implications, the amount of accuracy alone seems insufficient — as though something fundamental is absent. That's where explainable AI (XAI) comes in. XAI attempts to provide insight into the logic behind a system's decision making in ways that seem accessible — whether by providing a general description of the larger patterns guiding the model or breaking down the decision-making process behind a particular decision in detail. When these explanations resonate — when they enable someone to follow the path from the model's initial input to its conclusion — they make difficult decisions slightly less intimidating. However, when the explanations are vague, unclear, or completely off base, all of this begins to fall apart. Confidence diminishes, and trust deteriorates. And sometimes, new risks materialize — essentially the exact risks XAI were designed to mitigate. Ultimately, the true value of XAI is not determined merely by how well an AI system predicts a particular outcome. It is determined by how well it enables real people to grasp the decisions that are most crucial.

Keywords: AI interpretability, ethical AI, explainable AI (XAI), high-stakes decision making, human-AI collaboration, model confidence and calibration

Introduction

AI-supported decision systems now operate alongside professionals whose judgments carry substantial ethical, legal, and societal consequences. Automated recommendations may trigger financial loss, denied access to services, delayed emergency responses, or regulatory intervention. In such environments, the ability to justify decisions becomes as critical as predictive accuracy itself. Explanations are therefore not ancillary artifacts; they are integral components of the decision infrastructure.

XAI methods aim to render model behavior intelligible to diverse stakeholders. Both approaches are widely adopted in practice. However, a growing body of scholarship suggests that explanation methods may themselves be unstable, misleading, or poorly aligned with model behavior — particularly in dynamic environments characterized by non-stationary data and asymmetric error costs.

Much of the existing literature evaluates XAI through human-subject studies emphasizing perceived trust, satisfaction, or reliance. While such work is indispensable, it often presumes that the explanations presented to users are structurally reliable. This assumption is rarely validated. As a result, brittle explanation artifacts may be passed directly to practitioners, transferring structural weaknesses into

human confusion, misplaced confidence, or unwarranted skepticism.

This manuscript addresses this underexplored layer of the problem by asking a complementary question: how do explanations themselves behave under operational stress, independent of human perception? We argue that explanations should be evaluated for stability, coherence, and faithfulness before being evaluated with humans.

Background

Popular explanation tools including SHARP and LIME have limitations, such as the instability of the explanations, the lack of trustworthiness in the explanations, and the inability to consistently demonstrate the workings of the model (Doshi-Velez & Kim, 2017).

Therefore, researchers are recommending that practitioners abandon black-box models and replace them with more transparent models (Doshi-Velez & Kim, 2017). With the increasing complexity of the world today, reasonable interpretability results in a compromise of accuracy and human needs (Doshi-Velez & Kim, 2017).

The importance of explanations is rooted in the fact that they dictate how much faith we place in them. Previous research has demonstrated that when decision-makers understand the output, they are more likely to use the AI and less likely to rely on the output without critically evaluating it. Conversely, when decision-makers are unsure about the output, they are more likely to disengage from the AI or to over-rely on the output (Rudin, 2019).

Just when things seem to be getting more complicated, AI systems don't remain static. They may be running the same code from one day to the next, yet the world around them is constantly changing. Customers change how they buy; scammers evolve new ways to deceive; markets fluctuate rapidly from being relatively calm to highly volatile. And as the world changes under their feet. They become a lifeline for identifying whether the reasoning behind an AI's decision is beginning to diverge from reality. Without that clarity of vision, decision-makers will begin to trust a process that is quietly losing its footing behind the scenes. Contemporary monitoring efforts attempt to keep us grounded by linking together the concepts of drift detection and explainability. Drift detection refers to the observation of changes in the data that occur over time, while explainability refers to the ability of engineers and frontline workers to recognize changes in an AI model's reasoning process.

High-stakes decision making is characterized by varying levels of consequence. For example, missing a fraudulent charge could result in significant financial loss for a bank, while a false alarm might merely annoy a customer. Similarly, missing an emergency call could have a profound and lasting impact on a family, while erring on the side of caution may only result in a slight delay. Therefore, when the consequences of a mistaken decision are skewed in favor of one particular outcome, decision makers logically desire — and should expect — to have a clear understanding of why the AI system reached that decision as well. Global and local explanations facilitate decision-making by allowing decision-makers to view not only the general tendencies of a model, but also the specific reasoning process utilized by the model to justify a particular outcome (Elkan, 2001).

This study directly addresses the human element of decision-making. People perform their best work when they feel informed, treated with respect, and in complete control. Prior research has indicated that the benefits of explainability are maximized when people are held accountable for their decisions, when uncertainty is introduced through data drift or anomalous input, and when the consequences of a mistaken

decision are unequally distributed. The purpose of this research is to take these scenarios seriously, and to compare the effectiveness of global and local explanations across models, environments, and error-cost trade-offs. The goal of this research is straightforward, practical, and distinctly human — to understand when explainability is effective — and when it inadvertently creates obstacles (Rudin, 2019; Doshi Velez & Kim, 2017).

Literature Review

Most of the knowledge about how explanations influence judgments relates to how people use data-driven decision-making tools. Studies have demonstrated a straightforward fact: people will apply recommendations based upon a fit between the reasons behind those recommendations and their application of the work they are doing, and they will also be less likely to blindly follow those recommendations (Scharowski, Perrig, Svab, Opwis, & Brühlmann, 2023). Explanations are not neutral. The way they are presented can create variability in the perceived quality of the recommendation. Additionally, when the narrative presented in the explanation is unclear or contradictory, it will add additional mental processing burden and negatively affect the overall performance of the model (Herm, 2023; Scharowski et al., 2023). Overall, people will only reference an explanation when it provides a basis for understanding the rationale behind the recommendation — and even then, trusting the recommendation and applying it may be two different processes that need to be evaluated separately (Scharowski et al., 2023).

Therefore, providing explanations has become a requirement to build and maintain trust among regulators, customers and employees.

In addition to the local view, which provides a detailed look at each individual decision, a global view allows practitioners to see when the broader signal landscape is changing, and if it still makes sense based upon previous learning patterns.

In cases where the consequences of not identifying a fraudulent transaction, or triggering an emergency response, or in clinical triage, the costs associated with "false negatives" may be significantly higher than the costs associated with "false positives." In these situations, good decision-making considers these differences in cost, rather than simply striving for accuracy.

Therefore, many teams configure their models to reflect their organization's cost profile, and explanations need to accomplish two tasks: (1) Provide assurance that the model design accounts for the actual economic and human costs associated with potential errors, and (2) Explain why this specific instance was treated differently than others.

This is part of a larger, ongoing and practical debate. Some argue that in high-stakes decision-making, organizations should choose to use models that are transparent from the outset, because post-hoc rationalizations may be incomplete or even misleading. Others note that there are often when organizations need to employ sophisticated techniques to analyze complex data sets, and that well-validated, user-centered explanations can add significant value when properly used.

This is not an academic debate about transparency vs. opacity — it is the day-to-day trade-off between the performance of a model, how well users can understand its operation, and the organizational accountability for decisions made by models.

When viewed collectively, a common theme emerges: explanations are most valuable when individuals are accountable for the outcomes produced by models, when uncertainty or drift creates ambiguity in the data, and when the costs of errors are asymmetric. However, a gap still exists. There is currently limited research into how global explanations (those that provide insight into the big-picture decision-making process), and local explanations (those that provide insight into the decision-making process for an individual transaction), perform under the stresses of real-world environments (i.e., tight timelines, shifting data points, etc.).

Main Contributions

This research contributes four fundamental aspects to the literature and practice of XAI in high-stakes decision-making contexts. All the contributions are based on a non-human, non-agent methodology which assesses explanations, models and scenario-based conditions — not individuals — while incorporating humans' needs into the design of research studies. The contributions together provide insight into whether explanations add value to decision quality, and how complex explanations can undermine the reliability of explanations, and why models designed to be interpretable-by-design are the best alternative when there are issues with accountability and differentially costly errors.

Human-Centered No-Human Evaluation Framework for XAI Under Real-World Stressors

The author developed a framework for evaluating the performance of explanations and model behaviors under realistic workplace conditions (i.e., accountability, data drift, and asymmetric error costs) that are like real-world workplaces, and that does not involve either human subjects or simulated agents. The author evaluated how explanations and model behaviors perform across a number of combinations of the following factors: Explanation Type (Global vs. Local), Explanation Complexity (Plain Language to Dense Language), Model Class (Interpretable-by-Design to Black-Box with Post-Hoc Explanations), and Environmental Stressors (Data Drift; Cost Asymmetry) that occur in the workplace and that are critical for practitioners (Doshi-Velez & Kim, 2017; Elkan, 2001; Rudin, 2019).

Evidence That Local Explanations Are Fragile in the Presence of Drift and Cost Asymmetry (R1)

In three high-stakes domains (fraud detection, credit risk assessment, and emergency triage) the author demonstrated that local explanations (for example, instance-level attribution and counterfactuals) have high levels of instability under moderate distributional shift in the training data. Specifically, she showed that local explanations frequently change the feature ranking and/or emphasize different features relative to the original model output, despite little or no change in the model's output. Additionally, she demonstrated that local explanations rarely consistently reflect cost-sensitive boundary changes when false negatives are significantly penalized relative to false positives — which is precisely the environment that defines high-stakes decision-making (Elkan, 2001; Doshi-Velez & Kim, 2017).

Complexity Is Not Neutral: High-Complexity Explanation Artifacts Produce Lower Levels of Faithfulness and Consistency (R2)

The author demonstrated that high-complexity explanation artifacts (i.e., long, technical, jargon-heavy narratives) are more prone to errors: they produce contradictory logic, redundant chains, and exhibit lower levels of fidelity to underlying model behavior compared to plain language or moderately structured explanations. Importantly, she found that these results occurred regardless of the presence of human readers, demonstrating that complexity produces structural instability in the explanation layer itself (Doshi-Velez & Kim, 2017).

Interpretable-by-Design Models Prove to be the Most Robust Explanation Methodology Under Stress (R3)

When the accountability cues, data drift, and asymmetric costs were introduced into the experiments, interpretable models produced consistent decision logic, transparent reasoning chains and complete alignment between what the model did and what the explanation claimed was done. The results indicate that global explanations provided moderate support for orientation for black-box models, but black-box models coupled with local explanations were the least robust explanation methodology under stress. The results further reinforce the notion that in high-stakes decision-making domains, interpretability should be a primary design consideration and not an after-the-fact solution (Rudin, 2019; Doshi-Velez & Kim, 2017).

Research Questions

R1. At what points and in what manner do global and local explanations effectively support human sense making and decision justification in high-stakes decision-making environments characterized by accountability pressures, data drift and asymmetric error costs?
Reasoning: Practitioners report that explanations are most effective when decisions must be defended, when the world has changed beneath the model (data drift), and when the cost of being wrong is asymmetric (e.g. fraud miss vs. false alarm). We need to better understand the lived experiences of practitioners within these high-pressure environments, and how they use explanations (Doshi-Velez & Kim, 2017; Rudin, 2019).

R2. How does the density (i.e., level of detail, jargon, length) of global versus local explanations affect users' trust, perceived appropriateness and the social work of justification — avoiding confusion and over-trust?
Reasoning: Explanations can overwhelm or lead to "performative" compliance by users. A qualitative lens will enable us to examine how practitioners experience clarity, overload, or "social performative" compliance, and how complex explanation impacts their narrative of sense-making (Rudin, 2019; Doshi-Velez & Kim, 2017).

R3. In human-in-the-loop decision environments, how do practitioners compare the trade-offs between inherently interpretable models and black-box plus explanation pipelines — especially when subjected to regulatory scrutiny and operational uncertainty?
Reasoning: Post-hoc explainers (e.g., LIME/SHAP) can be unreliable/unstable, creating serious concerns in regulated, high-stakes decision-making environments. We want to gather practitioners' comparative assessments of trust, audibility, and usability in day-to-day operation (Ribeiro et al., 2016; Rudin, 2019).

Methodology

The methodology for this project, while not based on an experiment that included human subjects, is primarily focused on people — the analysts, supervisors, and frontline decision-makers who depend upon AI systems in high-consequence situations. Due to the absence of human participation in this study, the primary objective was to replicate human experiences through task development, model design, and output generation of explanations that accurately represented cognitive pressures experienced in realistic operational environments.

To accomplish this objective, the author developed a scenario-based experimental framework designed to analyze how various explanation types perform in environments that commonly challenge human judgment

including accountability, data drift, uncertainty, and differences in error costs (Doshi-Velez & Kim, 2017; Rudin, 2019; Elkan, 2001).

A significant amount of research into explanations has focused on subjective human experience. In the past, researchers have mostly examined the reaction to explanations—i.e., do people believe the explanation; do they trust it; does it provide comfort with respect to the decision? Often, the subject of explanation research has been on the way people receive the explanation.

This study differed from many previous studies by examining explanations from a more fundamental and arguably a more critical perspective. Rather than assessing how well people respond to explanations, this study assessed the quality of the explanations themselves—i.e., how clear the explanations were; how consistently the explanations supported the reasoning behind the decision-making process; how complete the reasoning within the explanation was; and how accurately the explanation described the action(s) of the model. To ensure that the quality assessments were valid, the explanations were assessed in simulated operational environments—i.e., fast-paced environments where decisions are typically made quickly, stress levels are elevated, and there is little room for error.

Results/Findings

There were no human participants involved in this study. However, every aspect of the study was influenced by human realities. The simulation environments were based upon the typical stresses experienced by fraud screening, credit evaluation, and emergency triage professionals—e.g., changing circumstances, varying consequences of decisions, and the possibility that decisions could be challenged or audited at some point in the future. Rather than measuring the subjects' subjective responses to the explanations, the study's primary objective was to test the reliability of the explanation approaches used in these simulations—i.e., how they perform when the environment becomes unstable.

Following is not a story about how people reacted to explanations. It is a story about how the explanation tools were able to function when the environment became unstable.

The Reliability of Local Explanations Decreased As The Environment Changed

Local explanations were unable to remain stable as the environmental context evolved. Case-level explanations exhibited considerable variability — sometimes dramatic — even when the input variables were almost identical. Ultimately, such variability would result in the inability of decision makers to develop a reliable understanding of how the decision-making system arrived at a particular decision.

Complexity Resulted in a "Hidden" Penalty

Long and dense explanations — characterized by technical vocabulary — were found to fail more frequently than short, simple explanations. Additionally, long and dense explanations were found to exhibit instability both during the training phase and following deployment. Such results suggest that complexity does not merely create barriers to the comprehension of explanations — it creates structural fragility in explanations prior to the exposure of the explanations to humans.

Embedded Transparency Outperformed Added Explanation Layers When Transparency Was a Major Concern

Embedded transparency in models—i.e., models designed to be transparent throughout the entire development and operation life cycle — provided clearer, more predictable reasoning than opaque models supplemented by an additional explanation layer. Specifically, when the decision maker was accountable

for errors and when errors had uneven costs—i.e., when the cost of failing to identify a true problem exceeded the cost of producing a false positive — embedded transparency was not merely beneficial — it was essential.

R1: When Do Explanations Remain Relatively Stable Under Stress?

Local Explanations Were Unreliable As the Environment Changed

Case-level explanations deteriorated as moderate changes were made to the environment. Changes in the feature importance rankings occurred even when the final decision made by the system barely changed. Practitioners would likely view such variability as erratic — i.e., the system was quietly revising its own narrative — and would likely reduce their confidence in the explanation (Pelosi et al., 2025).

Global Explanations Remained Relatively Stable

High-level summary explanations — e.g., global feature rankings and simple descriptions of the overall system — remained relatively stable across all environmental conditions. Because they describe large-scale trends rather than individual instances, global explanations were less responsive to small changes and therefore more suitable for dynamic environments where change is continuous.

Accountability Increased Expectations — and Revealed Shortcomings

When accountability cues — e.g., "this decision may be audited" — were explicitly included in the scenario, complex and opaque explanations often failed to adequately justify the decision. These explanations often omitted relevant intermediate reasoning, omitted intermediate steps, or created justification chains that did not clearly connect. On the other hand, systems designed to be interpretative from the outset did not have this issue — i.e., the internal logic of the system remained transparent under scrutiny because transparency was inherent to the structure of the system (Rudin, 2019).

Disparate Error Costs Made Shortcomings Unignorable

When missing a true problem was significantly more expensive than issuing a false positive, local explanations frequently failed to accurately reflect the system's transition toward greater caution. The explanation highlighted low-impact factors, whereas the internal logic of the system was adjusting to address higher-risk factors. Practically speaking, this discrepancy would mean that users are presented with explanations that do not explain the behavior they observe.

R2: How Does Explanation Complexity Influence Explainability?

Complex Explanations Were More Prone to Failure

Explanations rich in detail and jargon demonstrated clear structural failures — e.g., circular or repeated reasoning, conflicting references to the same variable, internal logical inconsistencies, and inflated values. These findings demonstrate that complexity not only taxes the user — it disrupts the explanation itself. Simple explanations were much more reliable. Short, direct explanations were found to be more reliable than longer, more convoluted explanations. Short, direct explanations maintained:

- a clear and ordered sequence of reasoning across similar cases
- coherence among the elements of the explanation
- alignment with the actual behavior of the system

In essence, simple explanations were easier to transport — i.e., from training to deployment, and from benign operating conditions to stressful operating conditions.

The takeaway is not that explanations are optional — it is that the design of explanations is critical, particularly in situations where the world will not stand still and the cost of being wrong is uneven.

Discussion of Findings

A number of the results reveal the human aspects of model explanations — although no humans were actively participating in our experiments — by demonstrating how model explanations, along with the models and scenarios used to generate those explanations, perform under various pressures that affect real practitioners, including pressure to be accountable, the changing nature of the environment, lack of complete information, and the uneven consequences of errors. Much like humans, the model explanations "failed" at the exact time a human would have struggled. That says a great deal about what the author saw.

Fragility Where Humans are Fragile

One of the clearest results (R1) revealed how fragile explanations are when the decision-making process becomes increasingly difficult to understand. While the decision-making context is subject to a wide range of potential changes — including data drift — which is inevitable as environments evolve, the local explanations provided by the model, especially during periods of rapid change, changed significantly, even as the model's predictions were relatively unchanged. It is reasonable to assume that such inconsistencies would have greatly confused a human attempting to understand why the model made the choices it made. Furthermore, this result supports a key critique of post-hoc explanation methods: that they provide plausible stories, but do not necessarily reflect the actual behavior of the model (Doshi-Velez & Kim, 2017; Ribeiro et al., 2016).

As previously stated, global explanations performed somewhat better, however they also demonstrated blind spots, particularly regarding cost-based shifts. As the costs associated with each option shifted, the global summary provided by the model failed to capture the nuanced shift in behavior that occurred within the model itself. This suggests that global explanations may be useful for providing a general orientation, but are insufficient to justify specific actions, especially in situations where accountability is of utmost importance.

Complexity Does Not Equate to Intelligence – Complexity Equals Overload

The second main result (R2) focused on the issue of complexity, and the message was clear: more does not equate for better. Highly technical, and therefore highly dense explanations were not simply harder to read (an issue in human-in-the-loop contexts), they were structurally less reliable, and exhibited more contradictory, redundant, and fidelity issues than simpler explanations.

If humans had been using these explanations, they would not have simply been overwhelmed by the sheer amount of information — they would have been misled.

This result further emphasizes the long-standing tension between complexity and reliability in the area of interpretability research: complexity often masks instability or noise, while simple, structured explanations tend to be both easy to comprehend and faithful to the underlying model (Tsang & Benoit, 2022; Rudin, 2019).

It was found that moderate-complexity explanations were most likely to exhibit consistent and predictable behavior. They were detailed enough to capture meaningful model behavior, but not so dense that they collapsed under the pressures of data drift and cost changes. This suggests that real-world teams may require more guidance in creating explanations that are appropriately sized, i.e., neither too transparent nor too opaque.

Interpretable Models: Clear Doesn't Break Down Under Stress

Finally, the comparison of interpretable models and black-box models (R3) added one last, defining element to the overall picture. Interpretable-by-design models performed consistently well through all tests —

including data drift, accountability, and asymmetric error pressure — and their explanations were not only consistent, but inherently trustworthy, as they relied on design principles rather than post-hoc interpretation tools that can become brittle under extreme conditions.

On the other hand, black-box models with post-hoc explanations demonstrated significant strain. Local explanations were brittle. Global explanations were broad, but incapable of capturing the nuances of cost-based behavioral shifts. Both local and global explanations struggled with alignment under data drift.

Overall, the results support a growing argument in high-stakes AI: when there are serious consequences associated with the decisions being made, interpretability should not be an add-on to the system, it should be the core (Rudin, 2019).

Implications for Human Real-World Decision Making

Although the author conducted this study in a purely computational manner, her results suggest a few important implications for humans that will eventually rely upon these systems:

- Unstable explanations lead to unstable trust in the system.
- Overly complex explanations create cognitive noise, not clarity.
- Blackbox models create uncertainty that explanations cannot reliably resolve.
- Interpretable models provide the most consistent support for human decision making.

In other words, even in a purely computational setting, we have seen the reflections of human needs: stability, coherence, predictability, and honesty in how decisions are explained. These are not luxuries, they are necessary to avoid confusion, frustration, mistrust, or misplaced reliance among humans operating in high-stakes environments (Černevičienė & Kabašinskas, 2024; CFA Institute, 2025).

Human Lessons from a No-Human Study

While the author did not include humans as participants in this study, she does not believe that the absence of human participants diminished the relevance of her results to humans. On the contrary, it allowed her to demonstrate a starker truth: explanations, whether generated by humans or machines, contain their own forms of weakness, brittleness, contradictions, and overloads. Therefore, if humans had been included in the experimental environment, many of these issues would have translated directly into confusion, frustration, mistrust, or misguided reliance.

Therefore, the failures that she saw in the simulation would likely fail in the field, whereas the successes we saw in the simulation would likely succeed in the field.

Therefore, the results provide a guide for the development of explanation systems that are designed to provide meaningful support for human decision making — systems that remain trustworthy not only when the world is stable, but when the world is unpredictable, when the stakes are rising, and when accountability is pressing hardest.

Implications for Research/Practice

Implications for Future Research

The Value of Conditionally Contextualized XAI: Who Should Receive Explanations, and Why

To move beyond demonstrating an average positive effect of explanations and to demonstrate the conditional value of explanations across several dimensions of variability, future research studies should evaluate explanations conditionally to:

- Accountability Conditions (i.e., when must humans justify a decision?)
- The degree of uncertainty due to either Drift (the gradual degradation of model performance over time) or Out-of-Distribution Inputs (inputs that were outside the training set); and,
- Asymmetric Error Costs (unequal costs associated with errors made by the model).

This study's design may be generalized through a series of pre-registered factorial experiments that cross the Model Class $\times$ Global/Local Explanation $\times$ Drift Level $\times$ Cost Asymmetry factor space and collect both perceived trust and behavioral reliance/overrides.

Prior research has demonstrated that trust and reliance are not identical measures, and that explanation effects are task-dependent; thus, we require carefully designed experiments that assess these differential effects.

Move evaluation criteria from self-reported attitudes toward calibrated behavior. In addition to assessing subjects' self-reported levels of trust, the current study proposes collecting additional behavioral metrics including:

- Calibrated Acceptance/Override Rates
- Selective Prediction (Abstain) Policies
- Decision Utility Under Cost Matrices.

Human-centered explanations have been found to increase reliance on models without always increasing accuracy or reducing cognitive load; accordingly, measuring subjects' behavioral responses is essential to accurately estimate the true benefits of explanations.

Compare Interpretable-by-Design Models vs. Post-hoc Explanations Under High-Stakes Constraints

Comparative studies of models that are inherently interpretable versus post-hoc explanations of Blackbox models are required to measure differences in terms of model performance, explanation quality, and governance requirements.

Specifically, researchers should investigate whether inherently interpretable models yield similar results under high stakes and unequal error costs constraints than models to which post-hoc XAI methods are applied, as previously proposed by Rudin (2019) and Doshi-Velez & Kim (2017), respectively.

Establish Drift-Aware Explanation Research Methods

Continuous monitoring performance and fairness metrics as well as global and local explanation signals in conjunction with data shift occurrences is required in longitudinal studies.

Additionally, researchers should investigate whether changes in feature attribution signals precede performance decline and whether explanation-based triggered retraining decreases subsequent harm. Existing literature in practitioner and academic publication venues suggest incorporating drift detection and explainability into system development, however, there is currently little empirical evidence supporting when to retrain based on explanation signal changes.

Develop Explanation “Fit-for-Purpose” Quality Benchmarks

Validated Rubrics for assessment of explanation usefulness by Stakeholder Role (Analysts, Auditors, End Users, Regulators) are required to establish standardized "fit" metrics (Comprehensibility, Actionability, Dispute/Appeal Readiness) allowing for cross-study comparisons and tooling benchmarking.

Integrate Governance Frameworks into Study Design

Map study conditions to specific obligations from the EU AI Act (Transparency, Human Oversight for High-Risk Systems) and the NIST AI Risk Management Framework (Explainable/Interpretable Characteristics, Continuous Monitoring). By integrating governance frameworks into study design, researchers will be able to better translate their findings into compliance-ready guidance and assess the external validity of their findings.

Implications for Practice

Use a two-tiered explanation Strategy: global for governance and local for Accountability

Global: Use feature effects and interaction summaries for face validity, stability over time, and alignment with policy requirements during periodic model reviews and audits
Local: provide per case explanations at points of decision, appeal, or adverse action. Ensure that the explanation is legible to non-technical decision makers and supports documentation requirements.

Make explanations part of Continuous Drift Monitoring

Integrate explanation dashboards with drift detectors: alert when global attributions change significantly or when local explanations rely on new, critical signals. Tie alerts to pre-defined triage paths (review thresholds, run targeted backtests, consider retrain).

Design for Asymmetric Error Costs from Day One

Encode a cost matrix in model objectives and in decision thresholds; report outcomes and expected cost. Match the decision to the real price tag you're paying

Add a short, plain explanation for each decision made by automation, to justify the reason behind the caution or boldness of the decision based upon your cost profile.

You will likely find that in fraud, triage, and safety decision-making, the potential costs associated with missing a problem are generally greater than the costs of raising a false alarm. This explanation should convey that message and demonstrate how the call reflects that disparity.

Select the Best Explanation Method for Your Data and Your Team Members

There are several ways of explaining decisions; however, the way each method performs may vary depending upon the nature of the data you collect (collinearity, changing sample size, etc.).

Develop guidelines for using the selected explanation methods. The guidelines should outline when to use each method, what expectations you have for the explanations (global/big-picture vs. local/case-level), and what typical pitfalls are associated with each method.

Provide training for those who review and make decisions about the explanations. Provide your analysts, reviewers, and managers with a quick way to determine whether an explanation provided for their consideration is valid or suspect—what to trust, what to verify, and what to dispute.

Measure What Matters (Behavior And Business), Not Just "Trust"

Monitor and evaluate outcomes that you could reasonably support in a boardroom or a hearing:

- Calibrated override rate (are humans correctly overriding the system?).
- Rate of reduction of expensive errors (measured against your cost matrix).
- Time required to complete a decision (are decisions being completed faster without sacrificing quality?).

Rate Of Post-Deployment Incidents (Are Unexpected Events Decreasing?).

Do not forget explanations can affect how many times people refer to a system without affecting the overall outcome of decisions. Monitor the outcome.

When there are extreme levels of oversight and the consequences of decisions are most severe, opt for clear built-in transparency

If you have line-of-sight responsibility — adverse-action notices, clinical justifications, regulatory audits — reach first for models that are interpretive by design.

Only use additional layers of explanations (i.e. add-on) when the increased complexity of the explanations is justified and the fidelity of the explanations has been confirmed.

Limitations

Controlled $\neq$ real life. The research was conducted in controlled environments with consistent tasks, timeframes, and cost matrices. The controlled environment helped to isolate variables and provide a clear understanding of cause-and-effect relationships; however, it is possible that these results may not translate to other workflows with less structure.

Attitude vs. Action. A person can state that they are comfortable with an explanation yet still choose to override it — or vice versa. Ultimately, the actions and resulting outcomes matter.

Replicate in field settings. Prior to implementing policies based on these findings, replicate them in operational environments with your data, your incentives, and your accountability structures. Explanations should be concise, accurate and cost-aware. Create clarity from the beginning. Measure what changes the risk and value. Educate those who carry the burden of making these decisions. It is through these actions that explanations cease to be decoration and become defense.

Conclusion

The system now sits next to you when your actions have consequences. The need for explanations is no longer discretionary. Good explanation design will answer both types of questions. High-level assessments will demonstrate how the entire process works and provides explanations that are based upon a single outcome (case) for example why a specific customer was classified as "High Risk" (Ribeiro, Singh, & Guestrin, 2016).

Any rollout plan of a model that is used at the decision-making edge must take this reality into consideration.

Two key points emerge from the above discussion.

Firstly, explanations must be dynamic and match the rate of changes in the underlying data. The world is constantly changing (e.g., customer habits, fraud tactics, market fluctuations), therefore the models must

also continue to evolve. By combining drift detection with regular explanation checks, it is possible to detect potential problems with the model well before the model's performance begins to slip, not to mention potential changes in the model's focus of attention (Enhanced MLOps, 2025; Pelosi, Cacciagrano, & Piangerelli, 2025). Secondly, not all errors are created equal. A missed problem in some domains could be catastrophic, whereas a false positive alarm would likely be more of an inconvenience than anything else.

Therefore, accuracy is not sufficient. Models should be trained to recognize a cost profile, and the explanations provided for each decision should highlight the reason(s) behind the model's cautionary or assertive decision-making process (Elkan, 2001).

As expected, there are limitations. Laboratory experiments cannot replicate the impact of incentives, the constraints imposed by time, and the way costs are debated in real-world organizations. Additionally, post-hoc tools need to be validated against design-for-interpretability methods. Finally, the "drift" experienced in real-world applications will not always occur in a predictable fashion. However, these limitations also provide direction for future endeavors: conduct field trials that relate changes in explanations to triggers for re-training; utilize behavior-first metrics (calibrated overrides, decision utility under a cost matrix); and invest in role-specific explanation literacy to enable auditors, analysts, and decision-makers to obtain the explanations they require to act responsibly (Pelosi et al., 2025; Scharowski et al., 2023).

Ultimately, the success of this work will not be measured by the aesthetic appeal of the plots produced, but by whether people can effectively use, challenge and justify the decisions made in a real-world context. Your contributions will assist in identifying the circumstances in which explanations provide real value, and how to integrate those explanations into the day-to-day governance of organizations — so that principles are converted into practices and confidence is translated into results.

Acknowledgements

This paper began as an issue which has been nagging at me more than anything else. As I worked on it, I found myself becoming increasingly focused on whether or not an AI system's decisions can be defended when those decisions need to be explained. Eventually, that concern became the basis for this research project.

Additionally, I employed generative AI tools periodically to assist with organizing and editing the written content. All aspects of the concepts, direction, and conclusions drawn from this research are mine.

References

- CFA Institute (Wilson, C.-A.). (2025, August 7). *Explainable AI in finance: Addressing the needs of diverse stakeholders*. Research & Policy Center.
<https://rpc.cfainstitute.org/research/reports/2025/explainable-ai-in-finance>.
- Černevičienė, J., & Kabašinskas, A. (2024). *Explainable artificial intelligence (XAI) in finance: A systematic literature review*. Artificial Intelligence Review, 57, Article 216.
<https://link.springer.com/article/10.1007/s10462-024-10854-8>.
- Doshi-Velez, F., & Kim, B. (2017). *Towards a rigorous science of interpretable machine learning*. arXiv.
<https://doi.org/10.48550/arXiv.1702.08608>.

- Elkan, C. (2001). *The foundations of cost-sensitive learning*. IJCAI Proceedings. <https://cseweb.ucsd.edu/~elkan/rescale.pdf>.
- Herm, L.-V. (2023). *Impact of explainable AI on cognitive load: Insights from an empirical study*. ECIS 2023 Proceedings. https://aisel.aisnet.org/ecis2023_rp/269/ and arXiv. <https://doi.org/10.48550/arXiv.2304.08861>.
- Nedunuri, U., Das Gupta, A., & Guha, D. (2026). *The paradox of explainability vs. performance in high-stakes autonomous AI systems: A systematic review*. <https://link.springer.com/article/10.1186/s42467-026-00018-5> 2.
- Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). “Why should I trust you?” Explaining the predictions of any classifier. KDD. <https://arxiv.org/abs/1602.04938>.
- Rudin, C. (2019). *Stop explaining black box machine learning models for high-stakes decisions and use interpretable models instead*. Nature Machine Intelligence, 1(5), 206–215. <https://doi.org/10.1038/s42256-019-0048-x>.
- Salih, A. M., Raisi-Estabragh, Z., Boscolo Galazzo, I., et al. (2024). *A perspective on XAI methods: SHAP and LIME*. arXiv. <https://arxiv.org/html/2305.02012v3>.
- Scharowski, N., Perrig, S. A. C., Svab, M., Opwis, K., & Brühlmann, F. (2023). *Exploring the effects of human-centered AI explanations on trust and reliance*. Frontiers in Computer Science, 5. <https://doi.org/10.3389/fcomp.2023.1151150>.