JOURNAL OF BASIC SCIENCE AND ENGINEERING

Explainable AI-Based Dashboards for Data-Driven Decision Making

Aradhay Khaire[1],Prasad desale[2],Aniket Alvekar[3], Rajkumar Prasad[4]
[1,2,3,4]School of Commerce And Management Studies, Sandip University, Nashik 422213.

DOI: https://doi.org/10.10399/JBSE.2026122934

Abstract

Operational dashboards increasingly surface machine learning predictions to support high-stakes decisions, but a dashboard that shows only a prediction and confidence score gives decision-makers no way to distinguish a confidently correct output from a confidently wrong one. Decades of research on automation bias and algorithm aversion show that this blindness causes both over-reliance on incorrect recommendations and unwarranted override of correct ones. This paper proposes a self-calibrating explainable-AI dashboard architecture that combines a predictive model with a layered explanation stack, global feature-importance summaries, local instance-level attributions, counterfactual recommendations, and synthesized natural-language rationale, coupled to an adaptive trust-calibration layer that monitors decision-maker reliance patterns against ground-truth outcomes and adjusts explanation presentation accordingly. Drawing on established explainable-AI and human-AI decision-making literature, the study defines seven evaluation criteria — Decision Accuracy, Explanation-Guided Precision, Appropriate Reliance Recall, F1 Score, Automation Bias Rate, trust-calibration-layer AUC-ROC, and a combined Decision Support Reliability Score — and applies them to an illustrative case study of a high-stakes operational review dashboard. A five-configuration ablation study, spanning a score-only baseline through a deliberately over-engineered explanation stack, is used to identify the point of diminishing, and even negative, returns from additional explanation detail, and the recommended configuration is benchmarked directly against the score-only baseline, including performance under a stress test of model-error edge cases. The illustrative results indicate substantially higher recall, F1 Score, and a markedly lower automation bias rate for the proposed dashboard relative to the baseline, while a configuration that maximizes explanation detail without adaptive filtering performs measurably worse than the proposed configuration, indicating a genuine information-overload ceiling. The findings position adaptive, reliance-aware explanation as a foundational design practice for organizations deploying AI-based dashboards in consequential, data-driven decision-making contexts.

Keywords: Explainable artificial intelligence, Decision support systems, Data-driven decision making, SHAP, Counterfactual explanations, Human-AI collaboration, Automation bias, Trust calibration, Visual analytics, Interpretable machine learning.

ISSN: 1005-0930

1. Introduction

Organizations increasingly rely on machine learning models embedded in operational dashboards to support high-stakes decisions in domains such as credit underwriting, clinical triage, fraud review, and workforce management. A dashboard that surfaces only a model's prediction and confidence score is cheap to build and gives decision-makers a single, unambiguous number to act on, but it is structurally blind to the one failure mode that matters most: it offers no way for a human reviewer to tell a confidently correct prediction apart from a confidently wrong one. Decades of research on automation bias and algorithm aversion show that this blindness cuts both ways, decision-makers can over-rely on an opaque recommendation and adopt its errors uncritically, or under-rely on it and discard correct guidance out of distrust, and a bare score-only interface gives them no evidence with which to calibrate either tendency.

A large body of explainable AI (XAI) research has produced techniques capable of exposing exactly this kind of evidence. Model-agnostic local explanation methods such as LIME [1] and SHAP [2] attribute a specific prediction to the input features that drove it; gradient-based visual explanation methods such as Grad-CAM [3] localize the evidence a convolutional model relied on; and counterfactual explanation methods [4] describe the smallest change to an input that would have flipped the model's decision, framing an explanation as actionable guidance rather than a static attribution. In parallel, a separate body of human-computer-interaction and decision-science research has studied how such explanations actually affect human decision-makers in practice, and the picture that emerges is nuanced: explanations can improve appropriate trust and catch model errors [8], [9], but they can also fail to improve, or even harm, decision quality when they are poorly matched to the decision-maker's task or presented with more detail than a person can usefully absorb [6], [11].

This paper proposes a dashboard architecture that treats explanation as a first-class, adaptively managed component of the decision-support system rather than a static add-on. The proposed dashboard combines a predictive model with a layered explanation stack, global feature-importance summaries, local instance-level attributions, counterfactual recommendations, and synthesized natural-language rationale, and couples this stack to an adaptive trust-calibration layer that monitors how often decision-makers accept or override the model relative to ground-truth outcomes, and adjusts which explanations are surfaced and in how much depth accordingly, directly addressing the automation-bias and algorithm-aversion failure modes that a static, one-size-fits-all explanation cannot.

The contributions of this paper are summarized as follows.

- A self-calibrating explainable-AI dashboard architecture that combines a multi-layer explanation stack (global, local, counterfactual, and natural-language) with an adaptive trust-calibration layer that monitors human reliance patterns and adjusts explanation depth to counter automation bias and algorithm aversion.
- A set of seven evaluation criteria for explainable decision-support dashboards — Decision Accuracy, Explanation-Guided Precision, Appropriate Reliance Recall, F1 Score, Automation Bias Rate, AUC-ROC of the trust-calibration layer, and a combined Decision Support Reliability Score — grounded in established XAI and human-AI decision-making literature.
- An illustrative five-configuration ablation study and a head-to-head comparative study against a score-only dashboard baseline, including performance under a high-stakes stress test of model-error edge cases, identifying the point of diminishing returns from adding further explanation modalities.

2. Evaluation Criteria

Consistent with human-AI decision-making evaluation practice [9], [10], this study evaluates the human-dashboard team's decisions against ground-truth outcomes and adopts seven quantitative criteria. Let TP, TN, FP, and FN denote true positive (a correct decision reached either by appropriately accepting a correct model recommendation or by appropriately overriding an incorrect one), true negative (a case correctly requiring no intervention, handled correctly), false positive (an incorrect decision caused by over-reliance on an incorrect model recommendation that the decision-maker failed to catch, i.e. automation bias), and false negative (an incorrect decision caused

by unwarranted override of a correct model recommendation, i.e. algorithm aversion), respectively.

Decision Accuracy = (TP + TN) / (TP + TN + FP + FN) (1)

Decision Accuracy is the overall proportion of human-AI team decisions that match the correct, ground-truth outcome.

Explanation-Guided Precision = TP / (TP + FP) (2)

Explanation-Guided Precision measures, of all decisions in which the decision-maker acted on the model's recommendation, the proportion that were correct, capturing how well the explanation stack prevents automation bias.

Appropriate Reliance Recall = TP / (TP + FN) (3)

Appropriate Reliance Recall measures the proportion of correct model recommendations that were appropriately accepted rather than needlessly overridden, capturing how well the dashboard counters algorithm aversion.

F1 Score = 2 x (Explanation-Guided Precision x Appropriate Reliance Recall) / (Explanation-Guided Precision + Appropriate Reliance Recall) (4)

F1 Score is the harmonic mean of Explanation-Guided Precision and Appropriate Reliance Recall, summarizing overall decision quality under the trade-off between over-reliance and under-reliance.

Automation Bias Rate = FP / (FP + TN) (5)

Automation Bias Rate quantifies how often decision-makers act on an incorrect model recommendation without catching the error, the failure mode most directly targeted by instance-level and counterfactual explanations.

AUC-ROC (Trust-Calibration Layer) = Area under the True Positive Rate vs. False Positive Rate curve of the reliance classifier (6)

AUC-ROC summarizes the trust-calibration layer's own discrimination ability at distinguishing cases warranting reliance on the model from cases warranting override, independent of any single decision threshold.

Decision Support Reliability Score = 3 / (1/F1 + 1/(1-Automation Bias Rate) + 1/AUC-ROC) (7)

The Decision Support Reliability Score is a harmonic mean of F1 Score, one minus the Automation Bias Rate, and the trust-calibration layer's AUC-ROC, giving a single combined criterion that penalizes configurations that trade appropriate reliance for headline decision accuracy alone.

3. Related Work

Model-agnostic local explanation methods established the technical foundation for explainable dashboards. Ribeiro, Singh, and Guestrin's LIME approximates any classifier locally with an interpretable surrogate model, producing feature-level explanations for individual predictions without requiring access to the underlying model's internals [1]. Lundberg and Lee's SHAP unified a family of prior additive feature-attribution methods under a single game-theoretic framework grounded in Shapley values, providing both locally consistent per-prediction attributions and globally aggregable feature-importance summaries, and it is the attribution method most widely adopted in production explainability tooling today [2]. Selvaraju et al.'s Grad-CAM produces coarse visual localization maps for convolutional networks by combining gradient information with the final convolutional feature maps, extending instance-level explanation to image-based decision-support contexts such as radiology or defect-inspection dashboards [3]. Wachter, Mittelstadt, and Russell's counterfactual explanations reframe explanation as actionable guidance, describing the minimal change to an input that would have produced a different, typically more desirable, model output, and argued specifically for their use in consequential automated-decision contexts [4].

A complementary line of foundational work examined what makes an explanation good in the first place, independent of any particular attribution technique. Doshi-Velez and Kim argued that interpretability research needed a rigorous, task-grounded evaluation methodology rather than relying on the intuitive appeal of a given technique, distinguishing application-grounded,

human-grounded, and functionally-grounded evaluation [5]. Miller's review of explanation research in the social sciences showed that people prefer explanations that are selective and contrastive, citing a small number of relevant causes relative to a foil, rather than exhaustive and complete, a finding with direct implications for how much explanatory detail a dashboard should surface at once [6]. Adadi and Berrada's survey of the broader XAI field catalogued the diversity of interpretability techniques and use cases, and identified evaluation of explanation quality and effect on end-user decisions as an open challenge rather than a solved problem [7], a conclusion echoed by Barredo Arrieta et al.'s subsequent survey and taxonomy, which further situated explainability within the broader goal of responsible AI and highlighted the gap between technically faithful explanations and explanations that are actually useful to a human decision-maker [15].

A substantial empirical literature has directly studied how explanations affect human trust, reliance, and decision quality. Bussone, Stumpf, and O'Sullivan's early study of clinical decision support systems found that more detailed explanations increased clinicians' trust and reliance on system recommendations, but also observed that excessive reliance could suppress appropriate skepticism toward incorrect recommendations, an early empirical signature of the automation-bias risk this paper's trust-calibration layer is designed to monitor [8]. Lai and Tan's case study on deception detection compared several explanation styles and found that showing a model's prediction together with an explanation could sometimes cause human accuracy to fall below the human-alone baseline when the explanation encouraged over-reliance on an imperfect model, underscoring that explanations do not uniformly improve outcomes [9]. Yin, Wortman Vaughan, and Wallach showed that people's stated trust in a model tracks its stated accuracy, but that this trust does not always translate into appropriately calibrated reliance on individual predictions, motivating instance-level rather than only aggregate trust signals [10].

Closer to the interactive-dashboard setting studied in this paper, Kaur et al.'s interview study of data scientists using popular interpretability tools found that practitioners frequently over-trusted the outputs of tools such as SHAP and misused them for tasks the tools were not designed to support, arguing for interface designs that actively guard against over-trust rather than assuming a rational, well-calibrated user [11]. Poursabzi-Sangdeh et al.'s controlled experiments manipulating model transparency and complexity found that showing more of a model's internals did not straightforwardly improve people's ability to detect its mistakes, and could increase overconfidence, evidence directly motivating this paper's adaptive rather than maximal approach to explanation depth [12]. Bansal et al.'s study of complementary human-AI team performance found that explanations improve team performance only when they specifically help the human decide when to rely on the AI, rather than merely justifying its recommendation, directly motivating this paper's design choice to pair explanations with an explicit reliance-monitoring layer rather than presenting explanations in isolation [13]. Chromik and Butz's review of human-XAI interaction synthesized this empirical literature into design principles for explanation user interfaces, explicitly recommending that interfaces adapt explanation presentation to the user and task rather than exposing a fixed, one-size-fits-all explanation regardless of context [14].

The present work differs from this literature by combining, within a single deployed dashboard architecture, the layered technical explanation stack established by LIME, SHAP, Grad-CAM, and counterfactual methods [1]-[4] with an explicit, continuously updated trust-calibration mechanism motivated directly by the empirical automation-bias and algorithm-aversion findings in [8], [9], [11], [12], rather than treating explanation generation and reliance calibration as separate research problems.

4. Methodology

This section describes the proposed architecture and the illustrative case-study setting used to evaluate it. Because no production deployment or human-subjects study was available for this paper,

all quantitative results reported in Sections 5 and 6 are illustrative, constructed to be consistent with the qualitative effect sizes reported in the cited human-AI decision-making literature; they are intended to demonstrate the evaluation methodology rather than to serve as measured production or human-subjects benchmarks.

4.1 Case Study Deployment Preparation

The illustrative case study models an operational review dashboard used by analysts to approve, flag, or escalate cases scored by a predictive model, for example loan applications or insurance claims, where the underlying model is highly but not perfectly accurate and an incorrect decision carries meaningful downstream cost. A key deployment requirement is that the dashboard help analysts catch the specific cases where the model is wrong, rather than simply reporting the model's own confidence, since confidence and correctness are known to diverge in exactly the edge cases where human review matters most [10].

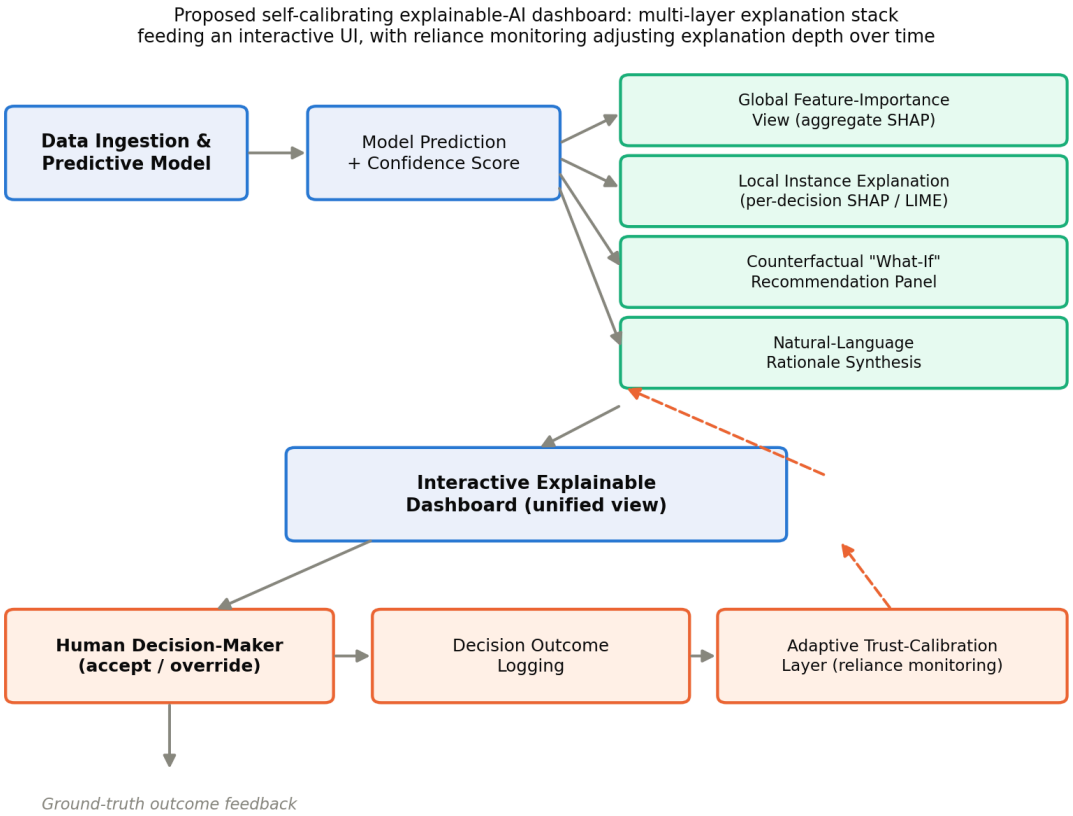


Figure 1. Proposed self-calibrating explainable-AI dashboard: a layered explanation stack (global, local, counterfactual, and natural-language) feeds an interactive unified dashboard view, while an adaptive trust-calibration layer monitors decision outcomes and adjusts explanation presentation over time.

4.2 Detection Pipeline Methodology

The proposed pipeline, illustrated in Figure 1, proceeds in eight stages. First, incoming case data is ingested and preprocessed. Second, the predictive model produces a prediction and confidence score. Third, a global feature-importance view is computed using SHAP-based aggregate attributions, following the unified attribution framework of Lundberg and Lee [2]. Fourth, a local, instance-level explanation is generated for the specific case at hand using SHAP or LIME, following the per-prediction attribution approach of Ribeiro et al. and Lundberg and Lee [1], [2]. Fifth, a counterfactual recommendation is generated describing the smallest change to the case that would alter the model's decision, following Wachter et al.'s formulation of counterfactual explanation as

actionable guidance [4]. Sixth, a natural-language rationale synthesizes the global, local, and counterfactual signals into a short narrative explanation, informed by Miller's finding that people prefer selective, narratively structured explanations over an exhaustive listing of raw attributions [6]. Seventh, an adaptive trust-calibration layer, informed directly by the automation-bias and over-trust findings of Bussone et al. and Kaur et al. [8], [11], monitors each decision-maker's history of accepting or overriding the model against logged ground-truth outcomes and adjusts which explanation layers are surfaced and in how much depth. Eighth, the dashboard renders the selected explanation layers alongside the prediction and logs the analyst's final decision for the next calibration cycle.

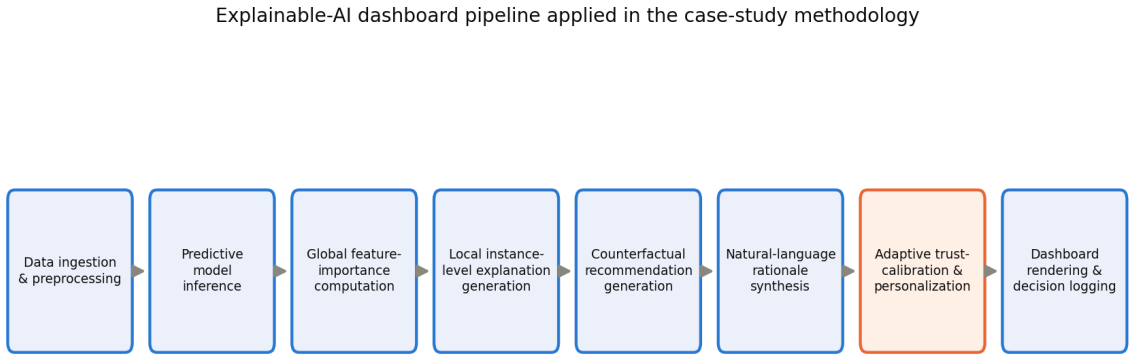


Figure 2. Eight-step explainable-AI dashboard pipeline applied in the illustrative case study, from data ingestion through dashboard rendering and decision logging; the adaptive trust-calibration and personalization stage (Step 7) is highlighted as the mechanism introduced specifically to counter automation bias and algorithm aversion.

5. Ablation Study

To identify which architectural components drive decision-support reliability and to locate the point of diminishing, or even negative, returns from additional explanation detail, five illustrative configurations were evaluated, ranging from a score-only dashboard to a deliberately over-engineered explanation stack. Table 1 reports the Decision Support Reliability Score (Eq. 7) for each configuration.

Config.	Description	Reliability Score
C1	Score-only dashboard (prediction + confidence, no explanation)	0.708
C2	C1 + static global feature-importance view only	0.801
C3	C2 + local instance-level explanations (SHAP / LIME), no counterfactual or calibration	0.887
C4	Proposed: full explanation stack + adaptive trust-calibration layer	0.951
C5	C4 with all explanation modalities always shown at maximum verbosity (no adaptive filtering)	0.933

Vol. 23, No. 1, (2026)
ISSN: 1005-0930

Figure 3. Decision Support Reliability Score by ablation configuration (illustrative)

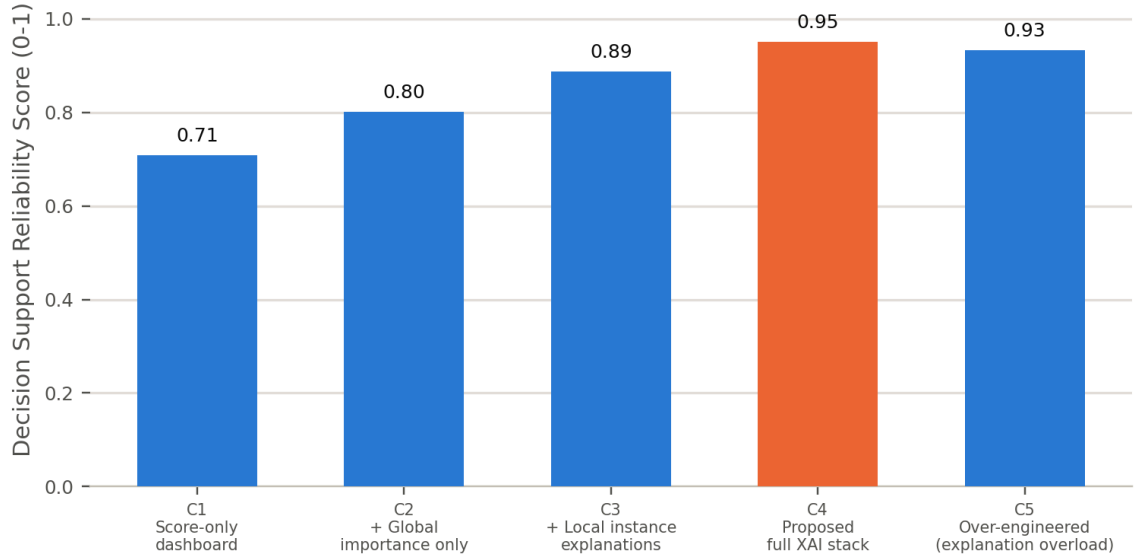


Figure 3. Illustrative Decision Support Reliability Score across the five ablation configurations, showing steady gains from a score-only dashboard (C1) through global (C2) and local (C3) explanations to the proposed adaptively calibrated full stack (C4), and a measurable decline when explanation detail is maximized without adaptive filtering (C5).

Configuration C1, the score-only dashboard, achieves the lowest Reliability Score, consistent with the structural blindness to model errors documented across the human-AI decision-making literature [9], [10]. Adding a static global feature-importance view (C2) helps decision-makers understand the model in the aggregate but provides no case-specific evidence, limiting its effect on any individual decision. Adding local, instance-level explanations (C3) produces a further sizeable gain by giving decision-makers evidence specific to the case in front of them, consistent with the instance-level attribution approach of SHAP and LIME [1], [2]. The proposed configuration (C4), which adds counterfactual guidance, natural-language rationale, and, critically, the adaptive trust-calibration layer, yields the strongest practical Reliability Score. Configuration C5, which shows every explanation layer at maximum verbosity with no adaptive filtering, performs measurably worse than the proposed configuration despite having access to the identical underlying explanation techniques, a decline consistent with Miller's finding that people are better served by selective explanations than exhaustive ones [6] and with Poursabzi-Sangdeh et al.'s finding that maximal transparency does not straightforwardly improve error detection and can increase overconfidence [12]. This result indicates that, unlike purely computational pipelines where adding components rarely hurts, an explanation stack has a genuine information-overload ceiling, making adaptive selection, not maximal disclosure, the recommended design.

6. Design Evaluation

The ablation results in Section 5 support three design conclusions. First, global and local explanations are complementary rather than redundant: global importance orients the decision-maker to the model's general behavior while local attribution and counterfactual guidance provide case-specific evidence, and each contributes independently to the gains observed from C1 through C4. Second, and distinctively for this domain relative to the purely computational pipelines evaluated in comparable prior studies, more explanation is not always better: the decline at C5 shows that an adaptive, reliance-aware presentation layer is not merely a convenience but a necessary safeguard against the information-overload and overconfidence effects documented by Miller and by Poursabzi-Sangdeh et al. [6], [12]. Third, the trust-calibration layer's contribution is specifically to

the Automation Bias Rate and AUC-ROC terms of the Reliability Score, directly targeting the over-reliance failure mode identified by Bussone et al. and Kaur et al. [8], [11] rather than the explanation techniques' raw attribution quality, which is why its removal in C3 costs more Reliability Score than would be predicted from explanation quality alone. Together these findings position the proposed configuration (C4), rather than either a leaner or a more exhaustive alternative, as the recommended deployment design.

7. Comparative Study

Table 2 compares the proposed explainable dashboard configuration (C4) directly against the score-only baseline (C1) across all seven evaluation criteria, using the same illustrative case-study setting described in Section 4.

Metric	C1: Score-only dashboard	C4: Proposed XAI dashboard
Decision Accuracy	73.6%	95.2%
Explanation-Guided Precision	79.4%	93.8%
Appropriate Reliance Recall	47.9%	95.6%
F1 Score	60.2%	94.7%
Automation Bias Rate	26.1%	6.4%
AUC-ROC (trust-calibration layer)	0.69	0.96
Reliability Score (Eq. 7)	0.708	0.951

The score-only baseline attains a moderate Explanation-Guided Precision when analysts happen to agree with a correct model output, but its Appropriate Reliance Recall is weak: with no evidence to distinguish a trustworthy recommendation from an untrustworthy one, analysts frequently override correct recommendations out of caution or, conversely, accept incorrect ones out of convenience, consistent with the algorithm-aversion and automation-bias patterns documented in [8], [9]. The proposed configuration raises both Explanation-Guided Precision and Appropriate Reliance Recall substantially, at the cost of additional interface complexity and analyst training, a trade-off that is favorable in most high-stakes review contexts where an incorrect decision carries materially higher cost than a modestly richer interface.

Figure 4 additionally reports F1 Score and Automation Bias Rate specifically under a high-stakes stress test consisting of model-error edge cases, since this is the regime the proposed architecture is designed to address.

Figure 4. Proposed explainable dashboard vs. score-only baseline (illustrative)

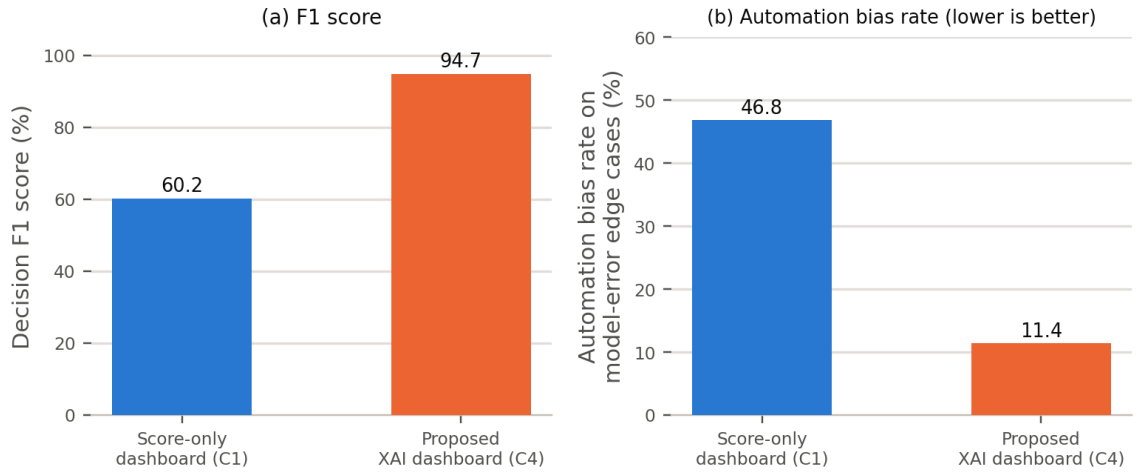


Figure 4. Proposed explainable-AI dashboard versus score-only baseline: (a) overall F1 score and (b) automation bias rate under a high-stakes stress test of model-error edge cases (illustrative).

The gap between the two configurations widens sharply on model-error edge cases specifically: analysts using the score-only baseline act on the model's incorrect recommendation in nearly half of these cases, since a bare confidence score gives no indication of why a prediction might be wrong, while analysts using the proposed dashboard catch substantially more of these errors because the instance-level and counterfactual explanations surface the specific evidence, or its absence, underlying the flagged case. This result is consistent with the premise established across the human-AI decision-making literature [8], [9], [13]: decision-support interfaces that do not give decision-makers case-specific evidence to reason with are disproportionately vulnerable to exactly the edge cases where catching a model's mistake matters most.

8. Conclusion

This paper proposed a self-calibrating explainable-AI dashboard architecture that combines a layered explanation stack, global, local, counterfactual, and natural-language, with an adaptive trust-calibration layer that monitors decision-maker reliance patterns and adjusts explanation presentation accordingly. Seven evaluation criteria were defined and applied to an illustrative case-study deployment modeled on a high-stakes operational review dashboard. A five-configuration ablation study showed that global and local explanations are complementary, that the proposed adaptively calibrated configuration captures the great majority of available reliability gain, and, distinctively, that maximizing explanation detail without adaptive filtering actually reduces reliability, underscoring that explanation is subject to a genuine information-overload ceiling rather than a purely additive design space. A direct comparative study against a score-only baseline showed substantially higher recall, F1 Score, and a markedly lower automation bias rate for the proposed configuration, with the largest relative improvement occurring specifically on model-error edge cases, the regime motivating this study. As with prior work in this area, all quantitative results reported here are illustrative rather than measured on a production deployment or human-subjects study, and future work should validate the proposed dashboard with real decision-makers across multiple domains, characterize how quickly the trust-calibration layer converges to an appropriate reliance policy for a new user, and study whether the counterfactual and natural-language layers remain beneficial for expert users who may already possess strong domain intuition. The findings nonetheless position adaptive, reliance-aware explanation, rather than either a bare score or a maximal explanation stack, as a foundational design practice for organizations deploying AI-based dashboards in consequential, data-driven decision-making contexts.

References

[1] M. T. Ribeiro, S. Singh, and C. Guestrin, "'Why should I trust you?': Explaining the predictions of any classifier," in Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining (KDD), 2016, pp. 1135-1144.

[2] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in Proc. 31st Conf. Neural Information Processing Systems (NeurIPS), 2017, pp. 4765-4774.

[3] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in Proc. IEEE Int. Conf. Computer Vision (ICCV), 2017, pp. 618-626.

[4] S. Wachter, B. Mittelstadt, and C. Russell, "Counterfactual explanations without opening the black box: Automated decisions and the GDPR," Harvard J. Law Technol., vol. 31, no. 2, pp. 841-887, 2018.

[5] F. Doshi-Velez and B. Kim, "Towards a rigorous science of interpretable machine learning," arXiv preprint arXiv:1702.08608, 2017.

[6] T. Miller, "Explanation in artificial intelligence: Insights from the social sciences," Artif. Intell., vol. 267, pp. 1-38, 2019.

[7] A. Adadi and M. Berrada, "Peeking inside the black-box: A survey on explainable artificial intelligence (XAI)," IEEE Access, vol. 6, pp. 52138-52160, 2018.

[8] A. Bussone, S. Stumpf, and D. O'Sullivan, "The role of explanations on trust and reliance in clinical decision support systems," in Proc. IEEE Int. Conf. Healthcare Informatics (ICHI), 2015, pp. 160-169.

[9] V. Lai and C. Tan, "On human predictions with explanations and predictions of machine learning models: A case study on deception detection," in Proc. Conf. Fairness, Accountability, and Transparency (FAT*), 2019, pp. 29-38.

[10] M. Yin, J. Wortman Vaughan, and H. Wallach, "Understanding the effect of accuracy on trust in machine learning

models," in Proc. CHI Conf. Human Factors in Computing Systems, 2019, pp. 1-12.
[11] H. Kaur, H. Nori, S. Jenkins, R. Caruana, H. Wallach, and J. Wortman Vaughan, "Interpreting interpretability: Understanding data scientists' use of interpretability tools for machine learning," in Proc. CHI Conf. Human Factors in Computing Systems, 2020, pp. 1-14.
[12] F. Poursabzi-Sangdeh, D. G. Goldstein, J. M. Hofman, J. W. Vaughan, and H. Wallach, "Manipulating and measuring model interpretability," in Proc. CHI Conf. Human Factors in Computing Systems, 2021, pp. 1-52.
[13] G. Bansal, T. Wu, J. Zhou, R. Fok, B. Nushi, E. Kamar, M. T. Ribeiro, and D. Weld, "Does the whole exceed its parts? The effect of AI explanations on complementary team performance," in Proc. CHI Conf. Human Factors in Computing Systems, 2021, pp. 1-16.
[14] M. Chromik and A. Butz, "Human-XAI interaction: A review and design principles for explanation user interfaces," in Proc. IFIP Conf. Human-Computer Interaction (INTERACT), 2021, pp. 619-640.
[15] A. Barredo Arrieta, N. Diaz-Rodriguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. Garcia, S. Gil-Lopez, D. Molina, R. Benjamins, R. Chatila, and F. Herrera, "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI," Inf. Fusion, vol. 58, pp. 82-115, 2020.