ETHIOPIAN INTERNATIONAL JOURNAL OF MULTIDISCIPLINARY RESEARCH

pISSN:2349-5707 Volume:13, Issue 02, February -2026 eISSN:2349-5715

Adaptive Explainable Artificial Intelligence for Governance-Oriented Risk Scoring in Organizational Change Management

Harrison D. Colebrook
University of Helsinki, Finland

ABSTRACT: Explainable Artificial Intelligence has progressively emerged as a foundational pillar in the governance of complex socio-technical systems, especially where algorithmic recommendations influence strategic organizational outcomes. Among the most sensitive of such environments is change management, in which decisions about system modifications, software updates, infrastructure reconfiguration, and operational process reengineering carry both financial and institutional risk. Contemporary organizations increasingly rely on Change Advisory Boards to evaluate, approve, or reject proposed changes, yet these decisions are now often informed by predictive models that estimate implementation risk. While such models promise improved accuracy and consistency, they also introduce epistemic opacity that can undermine trust, accountability, and regulatory compliance. This tension between predictive power and interpretability constitutes one of the most pressing challenges of modern AI-driven governance.

This article develops a comprehensive theoretical and methodological framework for integrating explainable artificial intelligence into predictive risk scoring for Change Advisory Board decision processes. Anchored in the emerging literature on algorithmic governance and model interpretability, the study positions explainable models not merely as technical artifacts but as institutional instruments that mediate between human judgment, regulatory requirements, and organizational legitimacy. Particular attention is given to recent advances in predictive risk scoring for change management, where machine learning models assess variables such as change scope, historical failure rates, interdependency structures, and operational volatility to generate probabilistic risk evaluations that guide CAB deliberations, as exemplified in the work of Varanasi (2025).

Drawing on a wide corpus of research on explainable artificial intelligence, interpretability metrics, counterfactual reasoning, feature attribution, and user trust, this article articulates how explanation methods transform opaque predictions into actionable and contestable knowledge. Through an extended conceptual methodology and a literature-grounded interpretive results section, the study demonstrates that explainable risk scoring enhances not only transparency but also procedural justice, stakeholder confidence, and long-term system resilience. The analysis further shows that explanation frameworks such as SHAP, rule ensembles, counterfactual profiles, and ceteris paribus plots can be aligned with governance principles in order to convert algorithmic outputs into decision-relevant narratives that Change Advisory Boards can evaluate, challenge, and refine.

The discussion situates explainable risk scoring within broader debates on algorithmic accountability, socio-technical trust, and the future of decision-support systems in organizational governance. It argues that without explainability, predictive risk systems risk becoming technocratic instruments that displace rather than support human judgment, whereas with well-designed explanatory mechanisms they can function as epistemic partners in collective decision making. By synthesizing engineering, management, and information governance perspectives, this article advances a model of explainable AI as an essential component of ethically and operationally sustainable change management.

Keywords: Explainable artificial intelligence; change management; predictive risk scoring; algorithmic governance; decision transparency; organizational trust

INTRODUCTION

The The rapid diffusion of artificial intelligence into organizational decision processes has fundamentally altered how institutions evaluate risk, allocate resources, and govern technological change. Nowhere is this transformation more consequential than in the domain of change management, where organizations must continuously modify their digital and operational infrastructures in response to evolving market conditions, cybersecurity threats, regulatory mandates, and innovation pressures. Traditionally, such changes have been evaluated by Change Advisory Boards, interdisciplinary committees tasked with assessing the potential risks, benefits, and organizational impacts of proposed modifications. These boards historically relied on expert judgment, precedent, and qualitative assessments, but the growing scale and complexity of modern digital systems have increasingly exceeded the capacity of purely human evaluation. As a result, machine learning models that predict the likelihood of failure, disruption, or cost overruns have been adopted to assist CAB deliberations, generating what is now often referred to as predictive risk scoring in change management (Varanasi, 2025).

While predictive models promise improved consistency and foresight, they also introduce a new epistemic problem: algorithmic opacity. Many of the most accurate machine learning models, including ensemble methods and deep neural architectures, operate as black boxes whose internal reasoning is inaccessible or unintelligible to human decision makers. This opacity is not merely a technical inconvenience; it is a governance challenge. Change Advisory Boards are accountable to regulators, customers, and internal stakeholders for the decisions they make, and a recommendation that cannot be explained cannot be defended, audited, or ethically justified (Olateju et al., 2024). The inability to articulate why a model assigns a high or low risk score to a proposed change undermines trust, limits institutional learning, and increases the likelihood of blind reliance on automated outputs.

The field of Explainable Artificial Intelligence has emerged precisely to address this tension between predictive performance and interpretability. XAI seeks to develop methods that render machine learning models transparent, intelligible, and contestable without necessarily sacrificing their accuracy (Adadi and Berrada, 2018). Over the past decade, a diverse ecosystem of explanation techniques has been developed, ranging from feature attribution methods such as SHAP to rule-based approximations, counterfactual explanations, and local surrogate models (Lundberg and Lee, 2017; Guidotti et al., 2018). These approaches aim to answer fundamental questions about AI systems: which factors most influenced a prediction, how a different input might have produced a different outcome, and what underlying patterns the model has learned from data.

Despite the growing sophistication of XAI techniques, their integration into organizational governance processes remains uneven and theoretically underdeveloped. Much of the existing literature has focused on technical benchmarks, visualization tools, or user studies in isolated domains such as image classification or medical diagnosis (Confalonieri et al., 2021; Hassija et al., 2023). Far less attention has been paid to how explanations function within institutional decision frameworks such as Change Advisory Boards, where algorithmic outputs are negotiated, challenged, and contextualized by multiple stakeholders with divergent interests. The work of Varanasi (2025) represents a significant advance in this respect by demonstrating how predictive risk scoring models can be embedded within CAB workflows, but even this contribution leaves open the deeper question of how explainability reshapes the epistemic authority of such systems.

From a theoretical standpoint, the introduction of explainable AI into change management can be understood through the lens of algorithmic governance. This perspective treats AI systems not merely as tools but as actors that participate in the production of organizational order by shaping what is seen as risky, acceptable, or optimal (Machlev et al., 2022). When a predictive model assigns a risk score to a proposed change, it effectively frames the decision space of the CAB, highlighting certain concerns while obscuring others. Explainability intervenes in this framing process by making visible the assumptions, correlations, and historical patterns that underlie the model’s judgment, thereby enabling human actors to interrogate and potentially revise those framings (Shin, 2021).

Historically, governance mechanisms in organizations have relied on procedural transparency to establish legitimacy. Decisions that affect multiple stakeholders are expected to be accompanied by reasons that can be communicated, debated, and, if necessary, contested. In the absence of such reasons, authority becomes arbitrary. Black-box AI threatens to erode this principle by introducing decision inputs that are epistemically opaque even to their designers (Guidotti et al., 2018). XAI can therefore be seen as a technological response to a normative requirement: the need for algorithmic systems to provide reasons that align with organizational standards of accountability and fairness (Rane et al., 2023).

The literature on trust in AI consistently shows that users are more likely to accept and rely on algorithmic recommendations when they can understand how those recommendations were produced (Bernardo and Seva, 2023; Yu and Li, 2022). In the context of change management, where decisions often involve trade-offs between stability and innovation, this trust is particularly crucial. A CAB that cannot explain why a change was approved or rejected risks internal resistance, reduced compliance, and even legal liability. By contrast, when risk scores are accompanied by intelligible explanations that link them to observable system characteristics and historical outcomes, decision makers can integrate them into a broader deliberative process rather than treating them as unquestionable outputs (Varanasi, 2025).

The problem this article addresses, therefore, is not merely how to make predictive models more transparent in a technical sense, but how to conceptualize explainability as an integral component of governance-oriented risk scoring. There remains a significant gap in the literature regarding how XAI methods can be systematically aligned with the institutional logic of Change Advisory Boards, which operate at the intersection of technical assessment, organizational politics, and regulatory oversight. While numerous studies have examined explainability in finance, healthcare, and consumer analytics (Ozkurt, 2024; Behera et al., 2023), the domain of change management has not yet received a similarly comprehensive theoretical treatment.

This article seeks to fill that gap by developing a detailed framework for explainable predictive risk scoring in CAB decision environments. Building on the insights of Varanasi (2025) and a broad range of XAI scholarship, it articulates how explanation techniques can be mapped onto the specific informational needs of change governance. The central argument is that explainable AI transforms predictive models from opaque oracles into epistemic partners that support deliberation, learning, and accountability. By making the logic of risk assessment visible and contestable, XAI enables CABs to exercise informed judgment rather than defer blindly to algorithmic authority.

The remainder of this article proceeds through a comprehensive methodological exposition, a literature-grounded interpretive results section, and an extended theoretical discussion. Throughout, every analytical claim is situated within the existing body of XAI and governance research, ensuring that the argument remains anchored in established scholarship while advancing a novel synthesis tailored to the context of organizational change management.

METHODOLOGY

The methodological orientation of this study is rooted in an interpretive, literature-driven research design that treats existing scholarly work as both empirical evidence and theoretical material for the construction of a comprehensive explanatory framework. Rather than conducting primary data collection, the study systematically integrates insights from explainable artificial intelligence research, organizational governance theory, and change management analytics to derive a coherent model of explainable predictive risk scoring.

This approach is particularly appropriate given the conceptual and infrastructural nature of the research problem, which concerns how algorithmic explanations mediate institutional decision processes rather than how individual users interact with a single system (Confalonieri et al., 2021).

The methodological foundation rests on three interlocking pillars. The first is the domain-specific grounding provided by predictive risk scoring in change management, as articulated by Varanasi (2025). This work offers a concrete reference point for how machine learning models are currently used to support Change Advisory Board decisions by estimating the likelihood of change failure based on historical data, system complexity, and contextual variables. The second pillar is the extensive body of research on explainable artificial intelligence, which provides a diverse repertoire of techniques for rendering such models intelligible (Adadi and Berrada, 2018; Hassija et al., 2023). The third pillar is the literature on trust, governance, and organizational decision making, which frames explainability as a normative and institutional requirement rather than merely a technical feature (Shin, 2021; Olateju et al., 2024).

The methodological process began with a conceptual mapping of the decision environment of a typical Change Advisory Board. CABs operate through a structured yet inherently interpretive process in which proposed changes are evaluated in terms of risk, impact, urgency, and strategic alignment. Predictive risk scoring models introduce a quantitative dimension to this process by assigning numerical or categorical risk levels to each proposal (Varanasi, 2025). However, these scores only become meaningful when they are embedded in a narrative that explains how they were generated and what they imply. The study therefore treats explanation as a form of knowledge translation that converts statistical patterns into decision-relevant insights (Guidotti et al., 2018).

Within this conceptual mapping, various classes of XAI methods were identified and analyzed in terms of their suitability for governance-oriented risk scoring. Feature attribution methods such as SHAP quantify the contribution of each input variable to a particular prediction, thereby allowing CAB members to see which aspects of a change request most influenced its risk score (Lundberg and Lee, 2017). Local surrogate models approximate the behavior of a complex model in the neighborhood of a specific instance, providing a simplified representation that can be more easily understood by non-experts (Guidotti et al., 2018). Counterfactual explanations describe how a change in input variables would alter the model's output, thus enabling decision makers to explore hypothetical scenarios such as whether reducing the scope of a change would lower its predicted risk (Lewis, 2013; Dhurandhar et al., 2018).

The methodological framework also incorporates global explanation techniques, which summarize the overall structure and behavior of a predictive model. Rule ensembles and partial dependence profiles reveal how risk scores vary as a function of key predictors across the entire dataset, offering CABs a macroscopic view of the model's logic (Friedman and Popescu, 2008; Apley and Zhu, 2020). Such global insights are essential for institutional learning, as they allow organizations to identify systemic risk factors and refine their change management policies accordingly (Varanasi, 2025).

To ensure analytical rigor, the study applies a form of triangulation across these different explanation modalities. Rather than privileging a single method, it examines how feature attributions, counterfactuals, and global profiles complement and constrain one another. For example, a SHAP analysis might indicate that system interdependencies strongly contribute to a high risk score, while a counterfactual explanation might show that decoupling certain modules would significantly reduce that risk. Together, these insights provide a richer and more actionable understanding than either could alone (Hassija et al., 2023).

The methodological design further acknowledges the limitations of explainable AI. Explanations are themselves models and therefore subject to approximation error, bias, and potential manipulation (Das and

Rad, 2020). The study therefore critically evaluates the epistemic status of different explanation techniques, drawing on the literature on interpretability metrics and validation (Carvalho et al., 2019). In the context of CAB decision making, an explanation must not only be mathematically faithful to the underlying model but also cognitively and institutionally meaningful. An overly complex explanation may satisfy technical accuracy but fail to support deliberation, whereas an oversimplified one may misrepresent the true drivers of risk (Shin, 2021).

The final methodological component involves synthesizing these insights into a coherent governance-oriented framework. This framework conceptualizes explainable predictive risk scoring as a layered process in which raw data are transformed into model outputs, which are then translated into explanations, which in turn are interpreted and acted upon by human decision makers. Each layer introduces both opportunities and constraints, and effective governance requires alignment across them (Olateju et al., 2024). By articulating this layered structure, the study provides a systematic basis for analyzing how explainability shapes CAB decisions in practice.

RESULTS

The interpretive results of this study emerge from the systematic integration of predictive risk scoring research with the extensive literature on explainable artificial intelligence and organizational trust. The central finding is that explainability fundamentally alters the epistemic role of AI in Change Advisory Board decision processes. Rather than functioning as an opaque authority that delivers unchallengeable risk scores, an explainable system becomes a dialogical participant whose judgments can be interrogated, contextualized, and, when necessary, revised by human actors (Varanasi, 2025; Shin, 2021).

One of the most significant outcomes concerns the relationship between feature attribution and perceived fairness. When a CAB is presented with a risk score for a proposed change, its members are naturally inclined to ask why that score was assigned. Feature attribution methods such as SHAP provide a structured answer by ranking the variables that contributed most to the prediction (Lundberg and Lee, 2017). In the context of change management, these variables might include the historical failure rate of similar changes, the number of dependent systems, the experience level of the implementation team, and the time window for deployment. When these contributions are made explicit, CAB members can evaluate whether the model's reasoning aligns with their own understanding of organizational risk, thereby enhancing procedural justice and trust (Behera et al., 2023).

A second result concerns the role of counterfactual explanations in strategic deliberation. Change management is inherently about choosing among alternative courses of action, and counterfactual explanations provide a natural bridge between predictive analytics and decision making. By indicating how a risk score would change if certain parameters were modified, such explanations enable CABs to explore design alternatives rather than merely accept or reject a proposal (Dhurandhar et al., 2018; Lewis, 2013). For example, if a model predicts high risk due to extensive system interdependencies, a counterfactual might show that modularizing the change would reduce that risk, thereby transforming a rejection into a conditional approval. This aligns closely with the iterative and negotiated nature of CAB processes described by Varanasi (2025).

The analysis also reveals that global explanation techniques support organizational learning beyond individual decisions. Partial dependence plots and rule ensembles expose patterns across the entire dataset, allowing organizations to identify recurring risk drivers and adjust their change management policies accordingly (Apley and Zhu, 2020; Friedman and Popescu, 2008). Over time, this can lead to the institutionalization of best practices, such as limiting the scope of changes during peak operational periods or investing in additional testing for highly interconnected systems. In this way, explainable risk scoring contributes not only to better immediate decisions but also to the evolution of more resilient organizational structures (Machlev et al., 2022).

Another key result pertains to the mitigation of over-reliance on automation. The literature on trust in AI warns that opaque systems can induce either blind faith or undue skepticism, both of which undermine effective decision making (Yu and Li, 2022; Bernardo and Seva, 2023). Explainable systems, by contrast, foster calibrated trust, in which users understand both the strengths and the limitations of the model. In CAB contexts, this means that members are more likely to treat risk scores as informative but not definitive, integrating them with qualitative insights and organizational knowledge (Varanasi, 2025).

Finally, the results highlight the importance of aligning explanation formats with the cognitive and institutional needs of decision makers. Visual saliency maps or high-dimensional feature vectors may be appropriate for data scientists, but CAB members require explanations that map onto their existing categories of risk, such as operational stability, regulatory compliance, and customer impact (Olateju et al., 2024). The most effective explainable systems are therefore those that translate technical outputs into governance-relevant narratives, a finding that resonates with broader research on human-centered XAI (Shin, 2021).

DISCUSSION

The theoretical implications of these results are far-reaching, as they situate explainable artificial intelligence at the heart of contemporary organizational governance. By rendering predictive risk scoring intelligible, XAI reshapes the distribution of epistemic authority between human decision makers and algorithmic systems. Rather than replacing human judgment, explainable models reconfigure it by providing new forms of evidence and reasoning that can be incorporated into deliberative processes (Varanasi, 2025).

From a socio-technical perspective, this reconfiguration can be understood as a shift from algorithmic automation to algorithmic augmentation. In automation, the goal is to remove human involvement from decision making, whereas in augmentation the goal is to enhance human capabilities by providing better information and analytical tools (Shin, 2021). Explainable risk scoring clearly belongs to the latter category, as it equips CABs with insights that would be difficult to derive from raw data while preserving the space for normative and contextual judgment.

The literature on explainable AI supports this interpretation by emphasizing that explanations are not merely technical artifacts but communicative acts that shape how users perceive and engage with AI systems (Adadi and Berrada, 2018; Hassija et al., 2023). In the CAB context, an explanation is a claim about why a change is risky or safe, and such claims are subject to the same scrutiny and debate as any other form of organizational reasoning. This transforms the model from a black box into a participant in a discursive process, aligning algorithmic governance with democratic and managerial norms of accountability (Olateju et al., 2024).

However, this transformation is not without its challenges. One potential limitation is the risk of explanation overload. If a model generates too many or too complex explanations, decision makers may become overwhelmed and revert to heuristic judgments, thereby negating the benefits of XAI (Das and Rad, 2020). Another concern is the possibility of strategic manipulation, where explanations are selectively presented or framed to justify predetermined outcomes. This underscores the need for robust information governance standards that ensure explanations are both accurate and fairly represented (Rane et al., 2023).

The comparison of scholarly viewpoints further reveals a tension between local and global explanations. Local explanations are highly useful for understanding individual decisions, but they may obscure systemic biases or structural risk factors that only become visible at the global level (Guidotti et al., 2018; Apley and Zhu, 2020). Effective CAB governance therefore requires a balanced integration of both, enabling boards to make informed judgments about specific changes while also monitoring the broader performance and fairness of their predictive systems (Varanasi, 2025).

Future research should explore how these explanation frameworks perform in real organizational settings, particularly in terms of their impact on decision quality, stakeholder trust, and long-term resilience. There is also a need to investigate how different cultural and regulatory contexts shape the reception of explainable AI, as governance norms vary widely across industries and regions (Machlev et al., 2022; Ozkurt, 2024).

CONCLUSION

This article has argued that explainable artificial intelligence is not a peripheral enhancement but a central requirement for the effective and legitimate use of predictive risk scoring in Change Advisory Board decision making. By integrating insights from XAI research, organizational governance theory, and the domain-specific work of Varanasi (2025), it has shown how explanations transform algorithmic outputs into actionable and accountable knowledge. In doing so, explainable risk scoring enables CABs to navigate the growing complexity of technological change while preserving the principles of transparency, trust, and informed judgment that underpin sustainable organizational governance.

REFERENCES

1. Adadi, A., and Berrada, M. Peeking inside the black box a survey on explainable artificial intelligence XAI. IEEE Access, 2018.
2. Rane, N., Choudhary, S., and Rane, J. Explainable artificial intelligence approaches for transparency and accountability in financial decision making. SSRN Electronic Journal, 2023.
3. Varanasi, S. R. AI for CAB Decisions Predictive Risk Scoring in Change Management. International Research Journal of Advanced Engineering and Technology, 2025.
4. Shin, D. The effects of explainability and causability on perception trust and acceptance implications for explainable AI. International Journal of Human Computer Studies, 2021.
5. Guidotti, R., Monreale, A., Ruggieri, S., et al. A survey of methods for explaining black box models. ACM Computing Surveys, 2018.
6. Apley, W., and Zhu, J. Visualizing the effects of predictor variables in black box supervised learning models. Journal of the Royal Statistical Society Series B, 2020.
7. Friedman, J. H., and Popescu, B. E. Predictive learning via rule ensembles. Annals of Applied Statistics, 2008.
8. Lundberg, S. M., and Lee, S. I. A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems, 2017.
9. Olateju, O. O., et al. Exploring the concept of explainable AI and developing information governance standards for enhancing trust and transparency in handling customer data. Journal of Engineering Research and Reports, 2024.
10. Bernardo, E., and Seva, R. Affective design analysis of explainable artificial intelligence a user centric perspective. Informatics, 2023.

11. Yu, L., and Li, Y. Artificial intelligence decision making transparency and employees trust. Behavioral Science, 2022.

12. Behera, R. K., Bala, P. K., and Rana, N. P. Creation of sustainable growth with explainable artificial intelligence an empirical insight from consumer packaged goods firms. Elsevier, 2023.

13. Confalonieri, R., Coba, L., Wagner, B., and Besold, T. R. A historical perspective of explainable artificial intelligence. Wiley Interdisciplinary Reviews Data Mining and Knowledge Discovery, 2021.

14. Hassija, V., Chamola, V., Mahapatra, A., et al. Interpreting black box models a review on explainable artificial intelligence. Cognitive Computation, 2023.

15. Carvalho, D. V., Pereira, E. M., and Cardoso, J. S. Machine learning interpretability a survey on methods and metrics. Electronics, 2019.

16. Machlev, R., et al. Explainable artificial intelligence techniques for energy and power systems. Energy and AI, 2022.

17. Dhurandhar, P., Chen, P. Y., Luss, R., et al. Explanations based on the missing. NeurIPS, 2018.

18. Lewis, D. Counterfactuals. John Wiley and Sons, 2013.

19. Ozkurt, C. Transparency in decision making the role of explainable AI in customer churn analysis. Information Technology in Economics and Business, 2024.

20. Das, S., and Rad, P. Opportunities and challenges in explainable artificial intelligence. arXiv, 2020.