INTERNATIONAL JOURNAL OF SOCIAL SCIENCES BULLETIN
ISSN: (E) 3007-1917 (P) 3007-1909
Volume 4, Issue 3, 2026

EXPLAINABLE ARTIFICIAL INTELLIGENCE AND MANAGERIAL DECISION QUALITY: TRUST, HUMAN OVERSIGHT, AND ACCOUNTABILITY IN HYBRID DECISION SYSTEMS

Zeeshan Zaib Khattak[1], **Hamza Iftikhar**[*2]

[1]*NUST Business School, National University of Sciences and Technology (NUST), Islamabad, Pakistan*
[*2] *Government and Public Policy, Jinnah School of Public Policy and Leadership, NUST, Islamabad, Pakistan*

[*2]hiftikhar@Jsppl.Nust.Edu.Pk

Corresponding Author: *
Hamza Iftikhar
DOI: https://doi.org/10.5281/zenodo.22339295

Received	Accepted	Published
17 January 2026	02 March 2026	16 March 2026

Abstract
Artificial intelligence is increasingly used to inform recruitment, credit, procurement, risk assessment, resource allocation, and strategic planning. The managerial usefulness of these systems depends not only on predictive performance but also on whether decision makers can understand, appropriately trust, and critically appraise their recommendations. This conceptual integrative review investigates how explainable artificial intelligence (XAI) affects managerial decision quality in hybrid human–ai systems. Building on work related to explainability, trust in automation, algorithmic aversion, and organisational decision structures, it separates explanation quality from explanation availability and distinguishes calibrated trust from generalised confidence. The paper contends that explanations can enhance decision quality by facilitating comprehension, enabling error detection, supporting justification, and promoting organisational learning. At the same time, explanations may also lead to cognitive overload, automation bias, superficial rationalisation, or misplaced confidence when they are technically complex, unstable, or disconnected from managerial tasks. Human oversight is therefore treated as an active organisational capability that involves authority, competence, time, and responsibility, rather than as mere nominal presence of a person in the decision loop. Five propositions link explanation quality, task expertise, decision stakes, calibrated trust, and accountability to decision quality. The paper develops a staged model of responsible managerial decision support and proposes a 2 × 2 experimental design comparing explanation conditions with human-review conditions. It advances management research by reframing XAI from a technical attribute to a governance mechanism and by showing that the goal is not maximal trust in AI but rather appropriate reliance under conditions of uncertainty and contestability.
Keywords: *explainable artificial intelligence; managerial decision-making; decision quality; trust in automation; human oversight; accountability*

1. Introduction

Organisations increasingly deploy artificial intelligence to forecast outcomes and propose actions. In practice, managers may receive algorithmic evaluations of employee potential, supplier risk, customer churn, project feasibility, fraud, creditworthiness, or the likelihood of operational disruption. Such systems can expand the set of options under consideration, handle high-dimensional data, and increase speed and consistency. However, managerial choices cannot be reduced to prediction alone (Abbas et al.,

2022; Al-Barrow et al., 2021). They also depend on interpretation, value judgments, taking responsibility for consequences, and communicating with relevant stakeholders.

The opacity of advanced models creates a core organisational challenge. A recommendation can be statistically sound yet difficult for a manager to understand or defend (Iqbal et al., 2024). When a system offers no accessible explanation of the factors driving its output, managers cannot spot obvious errors or justify decisions to employees, customers, regulators, or senior leaders. Explainable artificial intelligence has been introduced to make model behaviour—or individual outputs—more intelligible (Bushra et al., 2022). Rai (2020) characterises XAI as a shift from black-box to glass-box systems, underscoring the role of explanations for users and stakeholders.

However, the idea that increasing explainability automatically improves managerial decisions is overly simplistic. Explanations may be accurate but unintelligible, comprehensible but misleading, or persuasive without faithfully reflecting the model. Feature-importance displays can focus attention on influential variables while obscuring interactions, data limitations, or uncertainty (iqbal et al., 2025ab). Managers may accept an explanation because it reads like a plausible narrative, even when the underlying recommendation is wrong. As a result, explanations can either enable critical oversight or serve as tools for automation bias (KhaKtak et al., 2018a,b).

Work on trust adds further complexity to this relationship. Whether people trust AI depends on system reliability, transparency, information representation, task characteristics, and user experience (Glisson & Woolley, 2020). Too little trust can trigger algorithmic aversion, in which managers reject useful recommendations after noticing errors (Dietvorst et al., 2015; Khattak et al., 2019). Too much trust can lead to automation bias, where recommendations are accepted without sufficient scrutiny. The appropriate goal is calibrated trust: reliance that matches the system's actual competence for a specific task within a defined operational environment.

Organisational design also plays a role. Shrestha et al. (2019) distinguish full delegation, sequential human–ai decision arrangements, and aggregated human–ai structures. Each way of allocating decision authority creates distinct risks. Full delegation may be efficient for high-volume, low-ambiguity tasks, but it may be inappropriate for consequential decisions that involve contested values. Sequential arrangements can let the AI advise the manager or allow the manager to specify alternatives for the AI to evaluate. Aggregated structures combine independent judgments but require a rule to resolve disagreements (Ibrahim et al., 2024; Iqbal et al., 2024).

This paper brings these research directions together to address three questions. First, through what mechanisms does explainability affect the quality of managerial decisions? Second, how do expertise, decision stakes, and accountability shape managers' use of explanations? Third, what organisational design enables responsible human oversight? Decision quality is defined broadly to encompass accuracy, consistency, timeliness, ethical defensibility, stakeholder legitimacy, and the ability to revise decisions. This approach prevents a technically correct prediction from being treated as the complete measure of a managerial decision (Khattak et al., 2017).

2. Methodology

To sort out the existing research at the intersection of explainable artificial intelligence and managerial decision-making, this paper does not conduct new experiments or field research on its own. Still, it adopts an integrative literature review method to collect and summarise past research findings from five fields: management, information systems, psychology, human factors, and machine learning — that is, to gather relevant studies already published across these fields and then conduct a unified collation and analysis.

This cross-disciplinary, integrative design is perfectly suited to the needs of this study. The core of the research to be conducted is an exploration of the relationship between explainable artificial intelligence (XAI) and managerial decision-making. The concepts and

methods used are varied, ranging from controlled experiments and conceptual frameworks for theory building to technical analysis and empirical case evidence from enterprise organisations. For a research topic with such an extremely broad range of content types, the integrative literature review method is the most suitable approach to organise it.

As early as 2019, scholar Snyder proposed that when research in a field is scattered and unsystematic, an integrative review can derive usable theoretical models; while even earlier in 2005, another scholar, Torraco, also emphasised that conducting such a review must not merely list the found studies in a straightforward narrative, but must complete the work of critique and integration.

For this literature search, a set of core keywords was combined to screen eligible literature. These keywords are "explainable AI", "managerial decision-making", "trust in AI", "algorithm aversion", "automation bias", "human-ai collaboration", "human oversight", and "decision accountability", all centred on the research's core theme.

When screening the literature, we prioritise peer-reviewed, published outcomes—peer review means that other scholars have examined articles in the same field, so their quality is more assured. At the same time, we have incorporated several classic foundational studies to help readers first understand the basic logic of trust in automation and algorithmic judgment. All studies finally included in the collation are coded and marked according to seven unified dimensions; that is, the core information of each study is marked with a set of standards. These seven dimensions are: the purpose of proposing explanations, the attributes of explanations themselves, the characteristics of AI users, the structure of decisions, responses triggered by trust, accountability mechanisms, and the decision outcomes recorded in the studies.

This review adopts the concept-centric integrative method proposed by Webster & Watson (2002) to carry out the research. All the collected research evidence is categorised into four top-level analytical categories for cross-comparison. These four categories are informational-level benefits, behavioural-level risks, organisational-level conditions, and managerial-level safeguards.

It should be noted that this review does not claim to comprehensively cover all relevant studies in the large body of cross-disciplinary literature in this field, nor does it calculate a pooled effect size for the included studies. Its core outputs are an analytical framework for managers and a set of propositions for subsequent research to verify.

3. From prediction to managerial decision quality

An artificial intelligence model generates an estimate or classification using encoded objectives and available data. A managerial decision then integrates this output into a wider decision process. For instance, a turnover model may predict that an employee is likely to leave; however, the manager must still determine whether intervention is warranted, what form it should take, and whether the variables used are appropriate. Likewise, a procurement model may produce a supplier ranking, yet managers remain accountable for trading off price, resilience, environmental requirements, and strategic dependency.

This separation clarifies that predictive accuracy and decision quality are distinct. Even a more accurate model can lead to an inadequate decision if the objective is misstated, relevant context is missing, or the output is used outside its valid domain. In the other direction, a manager may legitimately decline a statistically compelling recommendation when it violates a legal constraint or conflicts with an organisational value that the model does not represent. As a result, high-quality hybrid decisions depend on both analytical capability and contextual judgment.

Jarrah (2018) describes human–ai collaboration as symbiotic when the machine's analytical capacity enhances human intuition and contextual knowledge. Shrestha et al. (2019) reached a similar conclusion, showing that the preferred organisational structure varies with interpretability, search-space specificity, decision speed, and replicability. Taken together, these points suggest that explainable AI should be

assessed based on the managerial work it enables, rather than treated as an abstract attribute.

4. Explanation quality

4.1 availability, intelligibility, and fidelity

Explanation availability means that a system provides some account of its output. However, the quality of the explanation is multidimensional. To be suitable, an explanation should be intelligible to the intended user, directly relevant to the decision, sufficiently faithful to the system's modelled behaviour, stable when conditions are similar, and explicit about uncertainty. Highly technical explanations may meet developers' needs yet still fall short for operational managers. Conversely, simplified explanations may be easier to understand while concealing important limitations.

Miller (2019) contends that effective explanations are contrastive and socially situated: people commonly want to know why one outcome occurred rather than another, and what factors could have changed it. This is a managerial point. For example, a manager reviewing a rejected loan application may need to understand why the applicant was classified as high risk relative to an acceptable profile, and whether information that could be corrected would change the outcome. Broad model documentation can not substitute for explanations tailored to a specific decision.

Rai (2020) distinguishes interpretable models from post hoc techniques for explaining complex models. Post hoc explanations can still be valuable, but managers should not presume that a reasonable local explanation fully represents the underlying model. Accordingly, fidelity should be documented and aligned with the stakes of the decision. When consequences are severe, organisations may opt for simpler models whose reasoning can be assessed directly, even if a complex model yields only a marginal improvement in predictive accuracy.

4. E explanation as a cognitive resource

Explanations can improve decision-making through at least four mechanisms. To begin with, they enhance understanding by clarifying which information shaped a recommendation. Next, they make errors easier for managers to detect by identifying impossible values, missing context, or unsuitable proxy measures. Third, they help stakeholders by providing grounds for the recommendation. Finally, repeated exposure to explanations can promote organisational learning about recurring patterns and risks. These advantages are contingent on expertise that is relevant to the task. A subject-matter expert can assess whether influential variables are substantively reasonable. In contrast, a manager with limited knowledge may treat an explanation as evidence of validity. For that reason, explanations should be coupled with training on model limitations, base rates, uncertainty, and common failure modes.

Proposition 1: Explanation quality will be positively associated with managerial decision quality through improved comprehension and error detection, provided that managers possess sufficient domain and AI literacy.

5. Trust, reliance, and behavioural risks

5.1 calibrated trust

Trust in artificial intelligence should not be treated as an end in itself. Glisson and Woolley (2020) demonstrated that transparency and reliability shape cognitive trust, whereas representation and social cues may affect emotional reactions. In managerial contexts, generalised trust can be hazardous because AI competence is task-dependent: a model validated for one workforce or market may underperform in another. Accordingly, managers must develop a calibrated sense of when the system is likely to be beneficial and when escalation is necessary.

Dietvorst et al. (2015) reported that people may become reluctant to rely on algorithms after observing them commit errors, even when algorithms outperform human forecasters on average. Subsequent research indicates that granting users limited capacity to adjust forecasts can lessen this aversion (Diet Vorst et al., 2018). Together, these results endorse participatory control, although unconstrained modification can also undermine accuracy. The governance problem is to provide substantive authority while maintaining an auditable record of overrides and their outcomes.

Proposition 2: The relationship between explanations and decision quality is mediated by calibrated trust, such that both under-reliance and over-reliance reduce the benefits of XAI.

5.2 automation bias and persuasive explanations

Explanations can foster overconfidence when they make a recommendation seem internally consistent. Managers may equate ease of articulation with correctness and may also accept selective evidence that reinforces their existing views. When the system reports exact percentages without uncertainty intervals or clearly stated limitations, users may take away a level of certainty that is not justified. Likewise, risk categories that use colour-coding can streamline decision-making while obscuring borderline cases. The sequence of judgment is also consequential. If managers encounter the AI recommendation before forming an independent assessment, anchoring effects may narrow the scope of their evaluation. A different design would begin with an initial human judgment, then present the AI output and its explanation, and finally require a documented reconciliation. This approach makes disagreement easier to identify and limits passive acceptance, although it may increase the time required.

Proposition 3: explanations that are persuasive but omit uncertainty and failure conditions will increase confidence more than decision accuracy, thereby raising the risk of automation bias.

6. Human oversight as an organisational capability

Human oversight is frequently framed as keeping a person "in the loop," yet mere presence is not enough. Genuine oversight depends on four components: the authority to change or halt a decision, the competence to interpret the outputs, sufficient time and information to conduct a review, and accountability for the final action. Without these elements, a reviewer risks functioning as a ceremonial approver.

Organisations should separate routine reviews from exception reviews. Routine review is suitable when every case requires contextual judgment. Exception review may improve efficiency in high-volume processes when reliable triggers can flag low-confidence outputs, distributional shifts, protected-group disparities, or challenges raised by stakeholders. In either case, both approaches must include escalation procedures and auditable documentation.

Managerial accountability should not be shifted onto technology. When a manager accepts an AI recommendation, responsibility remains at the organisational level. Explanations can strengthen accountability by making the basis of a decision traceable, but they can also enable post hoc rationalisation. Accountability is meaningful only when someone can investigate the model, correct inputs, revise rules, and remediate harm.

Proposition 4: Human oversight will improve hybrid decision quality when reviewers have genuine authority, relevant competence, adequate time, and responsibility for documented reconciliation of human and AI judgments.

7. Decision stakes and governance proportionality

The appropriate level of explainability and review depends on the stakes involved, the reversibility of the outcomes, the degree of uncertainty, and the overall scale. For automated recommendations used in routine inventory replenishment, organisations may need monitoring and predefined exception thresholds rather than individualised, case-by-case explanations. By contrast, hiring, dismissal, credit denial, safety-related actions, and major investment decisions require more robust explanations and clear human accountability, since mistakes can be highly consequential and may be difficult to undo. High-stakes decisions also require balancing multiple values. A project-selection model can rank expected financial returns, but it may not capture equity, environmental impacts, or institutional commitments. Managers should therefore clarify which objectives are represented in the model and which fall outside it. Decision templates can be designed to require explicit consideration of the values that have been excluded, the

stakeholders who are affected, the uncertainty involved, and alternative courses of action.

Proposition 5: The positive effect of XAI on organisational legitimacy and decision quality will be stronger when explanation depth and review intensity are proportionate to decision stakes, uncertainty, and reversibility.

8. A staged model of responsible hybrid decision-making

The synthesis is grounded in a six-stage model. In the first stage, the organisation specifies the purpose of the decision, identifies legitimate objectives, and makes value judgments that cannot be delegated. In the second stage, it confirms that the data are relevant, that the model performs as required, that subgroup effects are appropriate, and that the system operates within established limits. In the third stage, it chooses an explanation format that fits the user's task and level of expertise. In the fourth stage, the manager records an initial or independent assessment whenever this is feasible. In the fifth stage, the manager evaluates the AI output, the provided explanation, the associated uncertainty, and any points of disagreement. In the sixth stage, the final decision, its rationale, any overrides, and the outcome are documented to support learning and auditing.

This model also holds that explanation occurs within a broader decision-making process. Even a technically strong explanation can not rectify invalid data or an illegitimate objective. Likewise, human review can not safeguard stakeholders if the reviewer is compelled to validate the system. Governance must therefore cover the entire sequence, from problem formulation to post-decision monitoring.

9. Proposed empirical study

The framework can be evaluated using a 2 × 2 scenario experiment. Managers or experienced professionals would be randomly assigned to receive an AI recommendation, with or without a decision-specific explanation and with or without a structured human-review protocol. The scenarios could cover recruitment, supplier selection, project funding, and credit approval. In selected cases, each scenario would contain an objectively assessable error or an omitted contextual factor.

The dependent variables would include decision accuracy, detection of the planted error, confidence, reliance on the AI recommendation, perceived accountability, and the quality of the written justification. Domain expertise, AI literacy, risk orientation, and prior experience would be included as covariates. The study design should determine whether explanations improve error detection or merely increase confidence. Follow-up qualitative debrief interviews could clarify how participants understood the explanations and resolved disagreements.

A field extension could compare teams that use AI decision support before and after the introduction of structured explanation and review protocols. Audit logs would enable the comparison of overrides and outcomes. Ethical approval and informed consent would be required, and organisational data should be anonymised. The empirical paper should avoid causal language if it relies solely on cross-sectional perceptions.

10. Discussion and implications

This paper presents three core contributions to the field of responsible AI and organisational decision-making.

First, it expands the evaluation criteria for decision quality beyond mere prediction accuracy, adding justifiability, modifiability, compliance, timeliness, and analytical support — all of which are indispensable conditions for maintaining trust and implementing accountability in high-stakes decision-making.

Second, it distinguishes explanation quality from explanation accessibility, preventing enterprises from treating a random dashboard listing feature-importance rankings as fulfilling all their obligations related to responsible AI.

Third, it redefines oversight as a capability that requires dedicated resources and formal authority to support, rather than a perfunctory symbolic check, thereby creating space for genuine audits and eliminating the need to perform superficial compliance routines solely to satisfy internal policies (Mahnoor et al, 2026).

For managers, the most critical actionable recommendation is to build the solution around their own decision-making neXAI, rather than following a generic technical checklist as a formality when procuring explainable artificial intelligence (XAI).

They must require suppliers to clearly demonstrate for whom each explanation is intended, what specific questions it aims to answer, how to assess the alignment between the explanation and the model's actual conditions, whether it clearly communicates the model's own uncertainty, and how to raise an objection if the system generates an error formally.

When conducting user training, training materials should include not only cases where the AI provided correct recommendations but also high-stakes failure cases caused by model-needs mismatches and uncalibrated models. This is to enable end-users to learn to rely on the system appropriately, neither unthinkingly following its outputs nor rejecting them outright from the start.

An enterprise's internal performance system must not only reward managers who act on the AI's recommendations. This incentive structure will undermine the genuine auditing of the system's outputs and fail to encourage careful, context-specific assessments.

Enterprises must also track patterns of disagreement between managers and the AI system over the long term. A consistently low rate of overriding the AI's recommendations may indicate two possibilities: either the model's outputs are truly excellent and reliable, or the team has fallen into automation bias or is afraid to challenge the system, hiding their inability to evaluate the system's recommendations critically. If the override rate remains consistently high, there are two possible explanations: either the model is misaligned with the enterprise's specific use case, or unresolved transparency issues have led to widespread distrust in the system's outputs.

Collecting and analysing the reasons for each override and its ultimate practical impacts can help the entire enterprise accumulate cross-team experience to apply to future AI deployments. The explanations provided to stakeholders affected by automated decisions must be tailored to their specific needs, rather than dumping a pile of jargon-heavy internal technical documents that non-professionals can not understand.

10.1 explanation as organisational communication

To frame explanations as internal organisational communication, the first thing to clarify is that people in different roles within a company have vastly different needs for explanations, and this alone generates many new specific requirements.

Model developers need access to details that can diagnose operational issues with the model, so they can trace information to identify which step went wrong and which segment of logic hit a snag. Business managers have different needs: they require justifications that can be used directly to make decisions, as well as a clear understanding of the model's own uncertainty—which judgments are made with high confidence and which remain unvalidated —so they do not make decisions in the dark. Auditors have yet another set of needs: they need records that can be traced back to their source, so they can sort out the context of every link, and obtain evidence that control measures have been properly implemented, before they can deliver a smooth audit conclusion. As for ordinary people affected by the decision, they neither understand technical parameters nor need to manage internal processes; they need clear, easy-to-understand decision bases, to know why they received a certain outcome, and have clear channels to apply for corrections, so they do not find themselves with no way to voice their concerns. It is simply impossible to meet all the needs of these four groups with a single, unified explanation template.

A layered explanation framework is far more effective. At the time a decision is made, the affected party is first given a concise, direct basis for the judgment, sufficient to complete the current assessment; if a retrospective review or detailed verification is needed later, a more in-depth, complete explanation can be pulled up at any time; and if a formal compliance audit process is required, full professional technical documents are available. The most critical baseline embedded in this layered design is that

these different levels of explanation must align consistently from start to finish: simplified, accessible explanations shared externally must never contradict the underlying model that supports the system's operation, and the model's true logic must not be distorted solely to make the explanation easier to understand.

Explanation also redistributes power dynamics within a company. If only technical experts can challenge the recommendations generated by the system, even if all relevant information is nominally made public, it will only further consolidate experts' power, failing to achieve the original goal of enabling more people to participate in oversight and dispersing power. To avoid this problem, enterprises must establish dedicated "translation" roles to convert technical jargon into plain language that ordinary people can understand, bridging the knowledge gap between different groups, and build the capacity for independent auditing, preventing experts'right to speak from becoming an unchallengeable absolute authority. Employees or customers affected by decisions can easily point out incorrect input data or request a re-review of decisions concerning them without requiring advanced statistical knowledge, avoiding being forced to suffer in silence simply because they cannot understand professional content. Whether a process is convenient and accessible is never an afterthought added at the last minute; it is a core standard for measuring the quality of explanations.

10.2 ethical quality and distribution of error

When evaluating the quality of an AI decision-making system, many people's first instinct is to calculate its overall accuracy—count how many of 100 decisions are correct and assume that the higher the number, the more reliable the system. However, relying solely on this single average value often hides the most easily overlooked equity issue: model errors experienced by different groups are distributed very unfairly, with some groups unlucky enough to encounter errors repeatedly, while others rarely make mistakes, and these disparities are all concealed by the high overall accuracy.

Take a practical example: a recruitment model used by an enterprise to screen resumes, when evaluated in aggregate, achieves higher overall predictive efficiency than previous purely manual screening, and most judgments are correct. However, for a specific minority group, the probability of being incorrectly judged as unqualified by the model, despite meeting the job requirements, is far higher than for other groups—that is, the false negative error of "rejecting a qualified candidate" occurs at a rate far above the average for this small group.

Therefore, when managers evaluate the decision quality of this system, they must not only draw conclusions based on aggregated metrics; they must disaggregate the data, separately verify the error distribution for each group, and check whether any group bears a disproportionate risk of errors. Many models today come with built-in explanation functions that can show the basis for their judgments, which may help identify whether the model overrelied on problematic proxy variables, that is, reference items that seem related to recruitment requirements but actually conceal bias. However, to truly close these loopholes of discriminatory treatment, the model's own explanations are far from sufficient. Specialised tests must be conducted for each group to verify differences in error rates repeatedly, and independent stakeholders with an interest in the decision must be involved in the audit to uncover hidden unfairness.

Beyond the fairness of the error distribution, another issue must be clarified: the costs of different errors vary widely, and the way an enterprise relies on model outputs must be adjusted to reflect these costs, rather than applying a single standard across all scenarios.

Whether it is misclassifying a legitimate transaction as fraudulent in a fraud detection scenario, rejecting a qualified job applicant in recruitment, incorrectly flagging an ordinary person as a high-risk individual in a public safety system, or barring an eligible applicant from accessing social welfare benefits—anywhere a decision model is used, two opposing types of errors will arise: one is "allowing an ineligible case to pass", and the other is "rejecting an eligible case". The impacts and losses caused by these two

errors when they fall on different people cannot be compared as equivalent.

Therefore, before enterprise leaders review any outputs from the model, they must first define clear red lines for acceptable errors: which types of errors must never be committed frequently, and which can have a slightly higher tolerance. If a trade-off between the two types of errors ultimately becomes necessary, the justifications for this choice must be documented in formal writing for retention, rather than being decided haphazardly.

The commonly discussed explainable artificial intelligence can indeed help sort out the context for these trade-offs, laying out the causes and consequences on the table to inform decision-making. Still, it cannot make core moral judgments for humans—it cannot rank the moral priority of harms caused by different errors. These types of judgments are fundamentally the responsibility of the enterprise's team, and the entire judgment process must be incorporated into formal regulatory procedures, allowing relevant stakeholders to raise objections, rather than being a black-box operation decided privately by a small group.

Limitations and future research

The framework proposed in this paper remains theoretical for now, and its implementation across various decision-making scenarios requires validation using real-world data. Simply put, it is currently only a logically derived set of ideas; whether it can be applied to specific decision scenarios such as loan approval, employee recruitment, and welfare eligibility assessment must be tested against user feedback and decision outcomes in real-world contexts.

The role that explanations can play may vary depending on culture, corporate hierarchy, industry, and regulatory environment. People in different regions inherently have different standards for what counts as a "clear explanation," and differences in organisational rank, occupation (such as being a doctor versus a teacher), and the strictness of the regulatory rules the industry must comply with can lead to the same explanation producing completely different effects for different people.

Tight deadlines prevent careful auditing of the model's recommendations; group discussions may make it easier to catch errors, but they may also amplify shared biases. If a pressing deadline leaves people overwhelmed, they have no capacity to think through potential issues when they receive the AI's conclusions and supporting explanations. They can only hastily sign off to approve the decision. If multiple people gather to review the conclusion, in the best-case scenario, someone may spot a logical loophole that others missed, identifying the model's flaws. Still, it is also possible that all participants already hold the same bias. After a round of discussion, that hidden bias becomes more entrenched, leading to even more serious errors.

Future research can be developed along several directions: comparing local explanations that target only individual cases versus global explanations that cover the full scope of scenarios, comparing explanations presented numerically versus those presented in narrative text, and comparing the two different sequences of reviewing the explanation before making a judgment versus making a judgment before reviewing the explanation.

Researchers must also distinguish between "whether users feel the explanation is clear enough " and "whether the explanation aligns with the model 's actual operation." These two standards are fundamentally distinct and must not be conflated: the former is the user's subjective perception after reviewing the explanation, while the latter is the explanation's objective accuracy.

Future research can also compare individual and team audits. Another research direction that can be explored in depth is: what different outcomes arise when a single person versus a group of people audits the same AI explanation.

Group audits should leverage each member's professional background to gain a more thorough understanding of the explanation. Still, corporate hierarchies may suppress dissenting opinions, leading to the same problem of over-reliance on artificial intelligence as individual audits. Team audits are inherently beneficial; people with different educational backgrounds and experiences can come together to uncover all the

issues in an explanation, but if there is a clear hierarchy within the team. Junior employees are afraid to voice opinions that differ from those of their leaders; the result is no different from an individual audit, or even worse. Everyone is afraid to question the AI's conclusions, leading to an even greater risk of over-reliance on automation than in individual audits.

12. Conclusion

Explainability can improve managerial decision-making, but its benefits are conditional. High-quality explanations support comprehension, error detection, justification, and learning; poor explanations can create an overload or false confidence. The objective is calibrated trust and accountable reliance, not universal acceptance of AI. Effective human oversight requires authority, expertise, time, and responsibility, while governance must be proportionate to the stakes and uncertainty. Xai should therefore be understood as one component of a responsible hybrid decision-making system rather than a technical solution for organisational accountability.

Declaration of generative AI and AI-assisted technologies

During the preparation and revision of this manuscript, the authors used ChatGPT and PaperpPl for language editing, grammatical correction, and improvements in clarity and readability. These tools were not used to generate or modify research data, conduct statistical analyses, fabricate references, or make substantive research decisions. The authors critically reviewed and verified all AI-assisted content and accept full responsibility for the accuracy, originality, and integrity of the published manuscript.

REFERENCES

Abbas, s., Al-Barrow, H., Abdullah, H. O., Alnoor, A., Khattak, Z. Z., & khaw, k. W. (2022). Encountering COVID-19, perceived stress, and the role of a health climate among medical workers. Current psychology, 41(12), 9109–9122. Https://doi.org/10.1007/s12144-021-01381-8

Al-barrow, H., al-matos, M., Alharbi, R. K., alnoor, a., abdullah, h. O., Abbas, s., & Khattak, Z. Z. (2021). Understanding employees' responses to the COVID-19 pandemic: the attractiveness of healthcare jobs. Global business and organisational excellence, 40(2), 19–33. Https://doi.org/10.1002/joe.22070

Bushra, m. F., ahmad, a., ahmad, w., khattak, z. Z., saeed, i., han, h., & ariza-montes, a. (2022). Empirical investigation of the domain knowledge and team advertising creativity typology: the case of nescafé coffee. Journal of retailing and consumer services, 69, 103086. Https://doi.org/10.1016/j.jretconser.2022.103086

Diet vorst, b. J. Simmons, J. P., & Massey, c. (2015). Algorithm aversion: people mistakenly avoid algorithms after seeing them make mistakes. *Journal of experimental psychology: general, 144*(1), 114–126. Https://doi.org/10.1037/xge0000033

Diet Vorst, b, J. Simmons, J. P., & Massey, c. (2018). Overcoming algorithm aversion: people will use imperfect algorithms if they can (even slightly) modify them. *Management science, 64*(3), 1155–1170. Https://doi.org/10.1287/mnsc.2016.2643

Glikson, e., & Woolley, A. W. (2020). Human trust in artificial intelligence: review of empirical research. *Academy of Management Annals, 14*(2), 627–660. Https://doi.org/10.5465/annals.2018.0057

Ibrahim, n., Rahim, Z. A., & Iqbal, M. S. (2024). Exploring generic green skills to enhance the TVET curriculum in Malaysia. International journal of academic research in business and social sciences, 14(10), 1518–1532.

Iqbal, M. S., rahim, Z. B. A., hussain, s. A., Ahmad, n., Kaidi, H. M., Ahmad, r., & Ziauddin, r. A. (2020). Mobile communication (2 G, 3G & 4 G) and the future interest in 5 G in Pakistan: a review. Indonesian Journal of Electrical Engineering and Computer Science, 15(3), 211–229.

Iqbal, M. S., rahim, Z. A., & ohshima, n. (2024). Triz-informed intervention framework for sustainable 4ir adoption in Malaysian manufacturing and related services (MRS) industries. *J Adv Res Appl Sci Eng Tech*, *50*(1), 203–19.

Iqbal, M. S., rahim, Z. A., & Omer Khel, q. (2025). Harnessing generative AI for educational innovation: a TRIZ-PLR perspective. Discover applied sciences, 7(7). Https://doi.org/10.1007/s42452-025-07467-3

Iqbal, M. S., & abdul rahim, z. (2025) b. Driving the adoption of emerging technologies in Malaysian micro-, small, and medium-sized enterprises: a strategic intervention framework. Journal of the international council for small business, 1–24. Https://doi.org/10.1080/26437015.2025.2539866

Jarrah, m. H. (2018). Artificial intelligence and the future of work: human-ai symbiosis in organisational decision making. *Business horizons*, *61*(4), 577–586. Https://doi.org/10.1016/j.bushor.2018.03.007

Khattak, Z. Z., Abbas, s., Khushnood, M., Kaleem, m., & Mufti, O. (2017). Effects of gendering on academic performance in Pakistani universities. *Journal of managerial sciences, 11*(3, special edition), 161–184.

Khattak, Z. Z., Faraz, m., Abbas, s., Manzoor, H., & Bangash, R. (2018a). Factors affecting working capital investment: an international evidence. *Abasyn Journal of Social Sciences, 11*(special issue: IGCETMA 2018), 115–127.

Khattak, Z. Z., Sallam, A., Abbas, s., & Khushnood, M. (2018b). Pyramidal ownership composition and the firm's capital structure policy. *Abasyn Journal of Social Sciences, 11*(special issue: IGCETMA 2018), 30–41.

Khattak, Z. Z., Abbas, s., & Kaleem, M. (2019). Servant leadership style and leadership effectiveness: the moderating role of the cognitive style index. *Global social sciences review, 4*(2), 165–172. Https://doi.org/10.31703/gssr.2019(iv-ii).22

Lee, J. D., & See, k. A. (2004). Trust in automation: designing for appropriate reliance. *Human factors*, *46*(1), 50–80. Https://doi.org/10.1518/hfes.46.1.50_30392

Miller, T. (2019). Explanation in artificial intelligence: insights from the social sciences. *Artificial intelligence, 267*, 1–38. Https://doi.org/10.1016/j.artint.2018.07.007

Rai, a. (2020). Explainable AI: from black box to glass box. *Journal of the academy of marketing science*, *48*(1), 137–141. Https://doi.org/10.1007/s11747-019-00710-5

Shrestha, Y. R., ben-menahem, s. M., & von Krogh, G. (2019). Organisational decision-making structures in the age of artificial intelligence. *California management review*, *61*(4), 66–83. Https://doi.org/10.1177/0008125619862257

Snyder, H. (2019). Literature review as a research methodology: an overview and guidelines. *Journal of Business Research, 104*, 333–339. Https://doi.org/10.1016/j.jbusres.2019.07.039

Torraco, R. J. (2005). Writing integrative literature reviews: guidelines and examples. *Human Resource Development Review*, *4*(3), 356–367. Https://doi.org/10.1177/1534484305278283

Webster, j., & watson, r. T. (2002). Analysing the past to prepare for the future: writing a literature review. *Mis Quarterly*, *26*(2), xiii–xxiii.

Mahnoor, A., Iqbal, s., & Iftikhar, H. (2026). Digital learning tools and social collaboration in classrooms: teachers'experiences in enhancing peer interaction. Viredaz do director, 23, e234854-e234854.