Algorithmic Explainability and Trust in AI-Assisted Workplace Decisions: A Quantitative Research Proposal

Document type	Research Proposal
Subject	Management and Artificial Intelligence
Academic level	Postgraduate
Referencing style	APA 7th edition
Approx. word count	925

Algorithmic Explainability and Trust in AI-Assisted Workplace Decisions: A Quantitative Research Proposal

Background and problem statement

Artificial intelligence is increasingly used to recommend, rank and support decisions in organisational settings. Yet technically accurate systems may still be rejected, over-trusted or used inconsistently if employees do not understand how outputs are produced. Human-centred explainability is therefore important not only for transparency but also for calibrated trust. Shin (2021) showed that explainability and causability can shape trust and acceptance, while Gkinko and Elbanna (2023) demonstrated that employee trust in workplace AI develops through cognitive, emotional and organisational processes. The problem is that many organisations invest in explainable interfaces without knowing whether perceived explainability actually produces trust that supports appropriate acceptance rather than blind reliance.

Aim and objectives

The proposed study aims to examine whether perceived algorithmic explainability increases employees' acceptance of AI-assisted workplace decisions through trust in AI. The objectives are to: (1) test the relationship between perceived explainability and trust in AI; (2) test whether trust predicts acceptance of AI-assisted decisions; (3) assess the direct relationship between explainability and acceptance; and (4) test the indirect effect of explainability on acceptance through trust.

Theoretical rationale

Trust is a central mechanism in human-AI interaction because users must decide whether to rely on system recommendations under uncertainty. The human-computer trust scale developed by Gulati, Sousa and Lamas (2019) provides a validated basis for measuring trust in technological systems. Explainability should support trust by making system reasoning more interpretable and enabling users to judge whether recommendations are plausible. However, the effect is unlikely to be purely informational. Leichtmann et al. (2023) found that explanations could improve decision performance and more appropriate trust in a high-risk task, while Wang and Ding (2024) showed that explanation rationality can affect behavioural trust and decision performance even when self-reported trust changes are less pronounced. This distinction supports examining both attitudinal acceptance and trust rather than assuming that an explanation automatically produces reliance.

Hypotheses

H1: Perceived algorithmic explainability is positively associated with trust in AI.

H2: Trust in AI is positively associated with employee acceptance of AI-assisted workplace decisions.

H3: Perceived algorithmic explainability is positively associated with employee acceptance of AI-assisted workplace decisions.

H4: Trust in AI mediates the relationship between perceived algorithmic explainability and employee acceptance of AI-assisted workplace decisions.

Methodology

A cross-sectional quantitative survey is proposed among employees who use, or have recent experience with, AI-supported decision tools at work. A purposive sampling strategy will be used because respondents must have sufficient exposure to evaluate the system. A target of at least 300 usable responses is appropriate for stable multivariate estimation and subgroup checks. Perceived explainability can be operationalised using items adapted from Shin (2021), while trust can be measured using the human-computer trust scale developed by Gulati et al. (2019). Acceptance should be measured as behavioural intention or willingness to use AI-assisted recommendations in relevant work tasks. All adapted items should undergo content review and pilot testing before the main survey.

Analysis plan

The measurement model will first be assessed for reliability and construct validity. Structural relationships can then be estimated using covariance-based SEM or PLS-SEM depending on distributional assumptions, measurement characteristics and the research purpose. The mediation hypothesis should be tested using bootstrapped indirect effects rather than the causal-steps approach. Relevant controls may include age, AI-use experience, managerial status and task criticality. Common-method risk should be addressed procedurally through item ordering, anonymity and construct separation, and statistically through appropriate diagnostic checks.

Ethics and contribution

Participants should receive clear information about voluntary participation, data use, anonymity and withdrawal. The study should not collect confidential employer data or proprietary AI outputs. The contribution is practical as well as theoretical: organisations need evidence on whether explanations generate informed trust rather than superficial reassurance. This is especially important because workplace studies show that trust in AI is shaped by user experience and organisational context (Gkinko

& Elbanna, 2023), while employee acceptance can depend on trust in the perceived capability of AI systems (Deriu, Pozharliev, & De Angelis, 2024). The proposed model therefore offers a focused test of explainability as an antecedent, trust as a mechanism and acceptance as an outcome.

Conceptual framework and measurement quality

The conceptual model treats explainability as an antecedent, trust as a mediator and acceptance as the outcome. This ordering should be reflected in the questionnaire design. Explainability items should refer to the employee's ability to understand the basis and logic of AI-supported recommendations rather than to technical transparency in the abstract. Trust items should capture confidence in the system's reliability and dependability, while acceptance should focus on willingness to use or rely on AI assistance within legitimate task boundaries. Because these constructs are conceptually related, discriminant validity is essential. A pilot sample should be used to identify overlapping wording, ambiguous items and ceiling effects before the main study.

Sampling and data-quality controls

Eligibility screening should confirm that respondents have actually used an AI-assisted tool at work; hypothetical users may evaluate AI differently from experienced users. The survey should include an attention check and a minimum completion-time rule, but exclusions should be specified before analysing hypotheses. If respondents are nested within organisations, standard errors may need adjustment for clustering. The study should also record the type of AI system and decision context because trust in a low-stakes writing assistant may not generalise to hiring, finance or healthcare decisions. These controls improve interpretation without unnecessarily expanding the core model.

Conclusion

This proposal positions trust as the psychological bridge between explainable AI and employee acceptance. By using validated measures and an explicit mediation model, the study can distinguish whether explainability merely improves perceptions of transparency or meaningfully supports appropriate adoption. The design also provides a foundation for later experimental or longitudinal work that tests how trust changes after repeated interaction with AI systems.

References

Deriu, V., Pozharliev, R., & De Angelis, M. (2024). How trust and attachment styles jointly shape job candidates' AI receptivity. Journal of Business Research, 179, 114717. https://doi.org/10.1016/j.jbusres.2024.114717

Gkinko, L., & Elbanna, A. (2023). Designing trust: The formation of employees' trust in conversational AI in the digital workplace. Journal of Business Research, 158, 113707. https://doi.org/10.1016/j.jbusres.2023.113707

Gulati, S., Sousa, S., & Lamas, D. (2019). Design, development and evaluation of a human-computer trust scale. Behaviour & Information Technology, 38(10), 1004-1015. https://doi.org/10.1080/0144929X.2019.1656779

Leichtmann, B., Humer, C., Hinterreiter, A., Streit, M., & Mara, M. (2023). Effects of explainable artificial intelligence on trust and human behavior in a high-risk decision task. Computers in Human Behavior, 139, 107539. https://doi.org/10.1016/j.chb.2022.107539

Shin, D. (2021). The effects of explainability and causability on perception, trust, and acceptance: Implications for explainable AI. International Journal of Human-Computer Studies, 146, 102551. https://doi.org/10.1016/j.ijhcs.2020.102551

Wang, P., & Ding, H. (2024). The rationality of explanation or human capacity? Understanding the impact of explainable artificial intelligence on human-AI trust and decision performance. Information Processing & Management, 61(4), 103732. https://doi.org/10.1016/j.ipm.2024.103732

Source note: References were selected from peer-reviewed academic journals and publisher records. The sample is written as an original illustrative example and does not reproduce article text.