ISSN: 2229-6956 (ONLINE)
DOI: 10.21917/ijsc.2026.0596

EVALUATING THE INFLUENCE OF ARTIFICIAL INTELLIGENCE FEEDBACK ON HUMAN CONFIDENCE AND DECISION ACCURACY IN COMPLEX PROBLEM-SOLVING TASKS

Omolola Mariam Oparinde[1], John Olalere Ogunlola[2], Oluwakemisola Adewole[3], Bisola Kafayat Oyebamiji[4], Dan Taiye Aremu[5] and Blessing Alice Alao-Olatunji[6]

[1,2,3,4,6]Research and Doctoral College, University of Greater Manchester, United Kingdom
[5]School of Computing and Digital Media, London Metropolitan University, United Kingdom

Abstract

The increasing use of artificial intelligence (AI) in decision-support systems has reshaped human performance in complex problem-solving, yet its effects on confidence, trust, and accuracy remain unclear. This study investigates how AI generated feedback influences decision accuracy, confidence, and reliance behaviour using 3,400 trial-level observations from 100 participants in a chess-based dataset. Results show that decision accuracy improves significantly after AI feedback (mean increase from 0.43 to 0.82), particularly under accurate feedback conditions. While misleading feedback reduces post-decision confidence, overall confidence shifts remain small (mean = 0.08). Reliance on AI is strongly associated with improved performance (r ≈0.66). Machine-learning models further confirm that trust behaviour and feedback accuracy, rather than confidence, are the strongest predictors of decision outcomes.

Keywords:

Artificial Intelligence, Calibration, Confidence, Decision Accuracy, Feedback, Human–AI Interaction.

1. INTRODUCTION

The growing adoption of artificial intelligence (AI) in decision-support systems has transformed the manner in which decisions are made in varying complex roles like healthcare, finance, engineering, and organisational management [1]. The modern AI-based decision-support systems (AIDSS) are not created to automate decisions but to support human cognition by offering recommendations, confidence clues and performance feedback [2]. The increasing popularity can be explained by the fact that AI systems are able to handle large amounts of information, identify latent patterns, and produce quick judgments that are potentially beyond the ability of human cognition [3][4].

Nevertheless, in spite of these benefits, human-AI collaboration creates novel cognitive and behavioural problems. Specifically, the AI feedback changes the classical decision-making processes by affecting the trust, dependency, and confidence calibration [5]. As shown empirically, human beings tend to follow the advice of AI, particularly when feedback is provided with the manifestation of confidence in it [6]. Nevertheless, this deference does not always lead to the performance improvement, it has its impact on the performance that is based on the accuracy of AI, the complexity of the task, and the framing of feedback [7]. Miscalibrated trust is likely to cause automation bias and over-reliance, or inappropriate dismissal of useful AI feedback in multi-step, complex, uncertaintyprone, and ambiguous problem solving situations [8].

1.1 RESEARCH AIM

This study sees to investigate the impact of AI generated feedback on the confidence of human decision-making, the accuracy of their decisions and the adaptive behaviour in complex multi-step problem-solving problems, as well as to establish the predictive models and behavioural profiles that can be interpreted to design the effective AI feedback mechanisms.

1.2 RESEARCH OBJECTIVES

- To evaluate whether AI feedback consistency (accurate vs inaccurate) modulates trial-level trust formation and reliance behaviour over time.
- To assess how AI feedback influences subsequent decision accuracy and self-confidence across repeated problem-solving trials.
- To develop machine learning models to predict confidence shifts and trust dynamics from gameplay and feedback patterns.
- To identify cognitive adaptation profiles (e.g., adaptive, resistant, overreliant) using unsupervised clustering and behavioural indicators.

This paper fills such gaps by finding out how AI feedback affects human confidence, trust dynamics, and accuracy of decisions on a fine-grained, trial-by-trial basis in complex problem-solving tasks. The study will determine predictive indicators of confidence transitions and specific cognitive adaptation profiles by taking experimental control of AI feedback reliability and explainable machine-learning models, and unsupervised clustering. With this, it provides evidence to inform the design of AI feedback processes that can maintain the accuracy of decisions and human confidence that is well-calibrated.

2. RELATED WORK

The interdependence but non-linearity between AI feedback and human confidence and the accuracy of decisions have been previously studied in empirical studies, which show these three constructs are interdependent but non-linear [9] [10]. As per the current literature, AI feedback is always capable of increasing the quality of human decisions in case it is credible, well-calibrated, and consistent with the knowledge of the domain [11]. In this situation, AI can be seen as a cognitive reinforcement mechanism, which increases user confidence and helps them make more decisive judgements [12]. Nevertheless, in cases of AI feedback

inconsistency, vagueness, or calibration to the processes of human cognition, the accuracy of a decision diminishes, which in many cases leads to sub-optimal decisions [13] [14] Such results suggest that AI feedback is not necessarily a constructive phenomenon: its usefulness is associated with context-specific variables and abilities of the user to process and believe the information presented [9]. A study on the effects of confidence and trust calibration also indicates that the AI feedback has a significant impact on human self-confidence without explicit outcome feedback [15]. Interference-level prompting devices, like confidence prompts and correction feedback were found to have a small positive effect on the accuracy of trust, but they do not always increase decision accuracy [16]. Also, continuously positive AI responses can overconflict user perception and build overtrust, and contrary responses can undermine trust and lead to indecisiveness or doubting [17][18]. Such ambivalent results indicate that there is always a conflict between confidence calibration and decision accuracy in AI-assisted judgement [8]. Longitudinal studies also support the idea that repeated exposure to valid AI feedback enhances dependence with time, usually at the expense of autonomous generative thought [19]. Reliance moderation and accuracy improvement at the long-term level can also be moderated by self-confidence calibration interventions; however, these findings are usually measured by aggregated or session level [20]. Therefore, the area of confidence dynamics on a trial level, temporary adaptation, and user heterogeneity has not been researched properly, especially in multi-step problem-solving situations [13]. This research paper tries to overcome these shortcomings by combining trial-level examination, forecasting modelling, and unpredictable grouping to move temporal confidence variations and particular cognitive adaptation patterns.

3. METHODOLOGY

3.1 RESEARCH PHILOSOPHY

A positive research philosophy is adopted in this study, on the premise that the relationships among AI feedback, human confidence, decision accuracy, and reliance behaviour can be measured objectively through numerical data and statistical analysis. A quantitative, experimental design is put to use to systematically investigate causal patterns in trial-level human–AI interaction data. A deductive and cross-sectional design is implemented, with secondary data being relied upon rather than primary data collection. This approach is deemed appropriate because the effect of accurate or misleading AI feedback on confidence, accuracy, and trust-related behaviour across repeated decision trials is being tested.

3.2 DATA CLEANING

A single dataset was created by merging the raw decision logs, and completeness was screened. Less than one percent of observations was affected when trials with missing confidence ratings, move data that were not present, or unusable engine evaluations were excluded from the dataset. Consistency was improved by standardising variable names, and derived measures such as confidence shift, accuracy deltas, and reliance were recalculated. No meaningful outliers or patterns of nonrandom missingness were found in the data.

3.3 VARIABLES AND OPERATIONAL DEFINITIONS

The main variables are defined as pre-feedback confidence (selfconf), post-feedback confidence (aiconf), pre-feedback accuracy, post-feedback accuracy, AI feedback correctness (feedback2), and reliance on the AI recommendation. To capture the behavioural change induced by feedback, derived variables for confidence change, accuracy change, and trust behaviour are computed. Central importance to the analysis is given to these derived variables because a summary of how participants adapt after exposure to AI advice is provided by them.

The key measures are defined as follows:

$$C_i^{\text{shift}} = C_i^{\text{post}} - C_i^{\text{pre}} \; N_\varepsilon(x_i) = \left\{ x_j : \left\| x_j - x_i \right\| \le \varepsilon \right\} \tag{1}$$

where C_i^{pre} is pre-feedback confidence and C_i^{post} is postfeedback confidence for trial *i*.

$$A_i^{\Delta} = A_i^{\text{post}} - A_i^{\text{pre}} \tag{2}$$

where A_i^{pre} and A_i^{post} denote pre- and post-feedback accuracy, respectively.

$$R_i = \begin{cases} 1, & \text{if the participant follows the AI recommendation,} \\ 0, & \text{otherwise.} \end{cases} \tag{3}$$

where R_i is the reliance index for trial i.

3.4 DATA PREPARATION AND PREPROCESSING

The raw decision logs were merged into a single analytical dataset and screened for missing values, unusable engine evaluations, and incomplete confidence ratings. Trials with incomplete information were removed, affecting less than one percent of observations. Categorical variables were encoded consistently, numerical variables were standardised, and derived measures such as confidence shift, accuracy delta, and reliance were recalculated to ensure comparability across trials. The purpose of preprocessing was to maintain analytical consistency before statistical modelling and machine-learning analysis. Standardisation was applied to numerical variables using:

$$z_i = \frac{x_i - \mu_x}{\sigma_x} \tag{4}$$

where x_i is the raw value, μ_x is the sample mean, and σ_x is the sample standard deviation.

3.5 DESCRIPTIVE AND INFERENTIAL STATISTICAL ANALYSIS

Exploratory analysis was conducted to describe the distribution of confidence, accuracy, and trust-related behaviours across the 3,400 trials. Group comparisons were made between accurate-feedback and misleading-feedback conditions using paired and independent t-tests, repeated-measures ANOVA, and correlation analysis. These tests were used to determine whether feedback correctness was associated with changes in confidence, decision accuracy, and reliance behaviour. The paired-sample t-test was defined as:

$$t = \frac{\bar{d}}{s_d / \sqrt{n}} \tag{5}$$

where $\bar{d}$ is the mean difference between paired observations, s_d is the standard deviation of the differences, and n is the number of paired trials.

The repeated-measures ANOVA framework can be expressed as:

$$F = \frac{MS_{\text{between}}}{MS_{\text{within}}} \tag{6}$$

where $MS_{between}$ is the mean square between conditions and MS_{within} is the mean square error.

Associations between variables were assessed using Pearson correlation:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{7}$$

This was used to examine relationships such as pre- and postconfidence, pre- and post-accuracy, reliance, and AI feedback correctness.

3.6 PREDICTIVE MODELLING

To supplement the hypothesis tests, supervised machinelearning models were trained to predict confidence shift, decision accuracy, and trust behaviour. For confidence shift, regression models such as Linear Regression, Random Forest Regressor, and Gradient Boosting Regressor were used. For decision accuracy and reliance behaviour, Logistic Regression, Random Forest, and Gradient Boosting classifiers were applied. Model performance was evaluated using RMSE for regression and accuracy and F1-score for classification.

3.7 CLUSTERING AND BEHAVIOURAL PROFILING

To identify heterogeneous adaptation patterns, unsupervised learning methods were applied to the most important behavioural variables. Specifically, K-Means and DBSCAN were used to detect latent user profiles, including adaptive, resistant, and over-reliant behaviours. This clustering step was intended to reveal behavioural subgroups that are not visible through aggregate statistical analysis alone.

The K-Means objective function is:

$$J = \sum_{k=1}^{K}\sum_{x_i \in S_k} \|x_i - \mu_k\|^2 \tag{8}$$

where K is the number of clusters, S_k is cluster k, x_i is an observation, and μ_k is the centroid of cluster k.

For density-based clustering, DBSCAN identifies clusters by comparing each point's neighbourhood size to a threshold:

$$N_\varepsilon(x_i) = \{x_j : \|x_j - x_i\| \le \varepsilon\} \tag{9}$$

where ε is the neighbourhood radius. Points with sufficiently dense neighbourhoods are grouped into clusters, while sparse points are treated as noise.

3.8 ETHICAL CONSIDERATIONS

Anonymised secondary data from the publicly available Mendeley Chess AI–Human Decision Dataset [22] are utilised in this study, and direct participant recruitment or intervention is not needed at any stage. Ethical compliance is kept in place through strict adherence to the dataset's licensing terms and by ensuring that no identifiable personal information is examined in the dataset or reported.

4. ANALYSIS

An open-source Mendeley dataset [22] comprising 100 participants and 3,400 valid decision trials from a chess-based AI-assisted problem-solving task was analysed in the study. In each trial, pre-feedback confidence, first-choice accuracy, AI feedback correctness, post-feedback confidence, post-feedback accuracy, and reliance behaviour were captured.

Table.1. Dataset variables

Variable	N	Mean	SD	Min	Max
Self-confidence (selfconf)	3400	0.47	0.23	0.00	1.00
AI confidence (aiconf)	3400	0.54	0.25	0.00	1.00
Pre-feedback accuracy	3400	0.44	0.50	0.00	1.00
Post-feedback accuracy	3400	0.72	0.45	0.00	1.00
Accuracy delta	3400	0.29	0.47	-1.00	1.00
Reliance index	3400	0.53	0.50	0.00	1.00

4.1 DESCRIPTIVE STATISTICS OF KEY VARIABLES

The Table.1 summarises the main behavioural variables used throughout the analysis. It supports the first objective by describing how confidence, accuracy, and reliance behaved across the full sample. Modest baseline confidence and comparatively low initial accuracy, followed by stronger post-feedback performance, are indicated by the descriptive profile. A mean self-confidence level of 0.47 was observed, and an increase to 0.54 after feedback was recorded. More importantly, a rise in mean accuracy from 0.44 to 0.72 was shown, while an average reliance index of 0.53 was observed, indicating that AI guidance was followed by participants in roughly half of the trials.

4.2 DISTRIBUTIONAL PATTERNS IN CONFIDENCE AND ACCURACY

A consistent pattern of upward movement after AI exposure is shown by the distribution plots in Fig.1–5. A broad spread in pre-decision confidence is indicated by Fig.1, while a slight rightward shift in post-feedback confidence is shown by Fig.2. A heavy concentration of pre-feedback accuracy toward incorrect responses is confirmed by Fig.3, with the difficulty of the task being reflected. By contrast, a marked improvement in post-feedback accuracy is shown by Fig.4. Additional confirmation that most trials improved after feedback, fewer remained

unchanged, and only a small fraction declined is provided by the accuracy-delta distribution in Fig.5.

Collective support for the first research objective is provided by these plots through the demonstration that measurable behavioural change at the trial level is associated with AI feedback.

4.3 EFFECTS OF FEEDBACK CORRECTNESS

To systematically investigate whether the quality of AI feedback made a meaningful difference, the dataset was split into accurate-feedback and misleading-feedback conditions. Clear differences between the two conditions are shown by the summary statistics in Table II. Higher post-feedback confidence, stronger post-feedback accuracy, and larger accuracy gains were linked to accurate feedback. Reduced final confidence and weaker performance improvement were produced by misleading feedback. The second objective is directly addressed by this pattern, and the conclusion that AI assistance is beneficial only when it is reliable is supported by it.

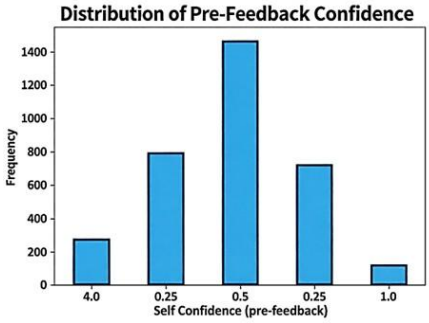

Fig.1. Pre-decision confidence

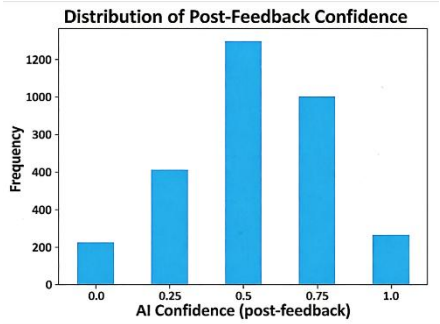

Fig.2. Post-feedback confidence

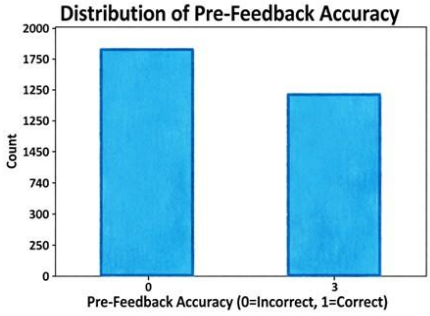

Fig.3. Pre-feedback accuracy

Consistent outperformance of the misleading-feedback condition by the accurate-feedback condition, especially in postfeedback accuracy and accuracy change, is made clearly visible by the visual comparisons. However, a weaker confidence pattern is observed between the conditions: although a higher final confidence level was produced by accurate feedback, only marginal differences between conditions were found in the average confidence shift. A stronger effect of AI feedback on outcome accuracy than on the magnitude of confidence adjustment is suggested by this.

Table.2. Mean Outcomes by AI Feedback Correctness

AI feedback condition	Self conf	Ai conf	pre_ accuracy	Post accuracy	Confidence shift	accuracy delta
Misleading feedback	0.42	0.49	0.28	0.50	0.07	0.22
Accurate feedback	0.53	0.59	0.62	1.00	0.06	0.38

4.4 CORRELATION STRUCTURE OF THE BEHAVIOURAL VARIABLES

A compact view of the dependency structure among confidence, accuracy, reliance, and feedback correctness is provided by the correlation heatmap in Fig.6. Pre- and post-confidence showed a weak positive relationship ($r \approx 0.25$), indicating moderate continuity in subjective assessment. Confidence shift was strongly related to post-confidence ($r \approx 0.65$) and negatively related to pre-confidence

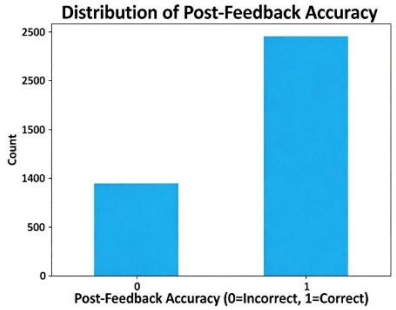

Fig.4. Post-feedback accuracy ($r \approx -0.58$), which is expected because confidence shift is a derived difference score.

Accuracy measures were more informative. Pre- and post-feedback accuracy correlated moderately ($r \approx .50$), suggesting that baseline ability remained relevant after AI input. Accuracy delta was negatively associated with pre-accuracy ($r \approx -0.57$) and positively associated with post-accuracy ($r \approx 0.42$), indicating that the largest gains occurred among participants who started from weaker baselines.

The most consequential relationships were observed for trust behaviour. The reliance index was strongly correlated with post-accuracy ($r \approx 0.66$) and accuracy delta ($r \approx 0.56$), showing that following AI guidance was associated with better decisions. Reliance had only a weak relationship with preaccuracy ($r \approx 0.06$), indicating that trust behaviour was not simply a function of prior skill. Positive associations of AI feedback correctness with post-accuracy ($r \approx 0.56$) and accuracy delta ($r \approx 0.17$) were observed,

and a positive correlation with reliance (r ≈ 0.27) was also found. Taken together, the heatmap suggests that performance gains were driven primarily by reliable feedback and appropriate reliance, not by confidence alone.

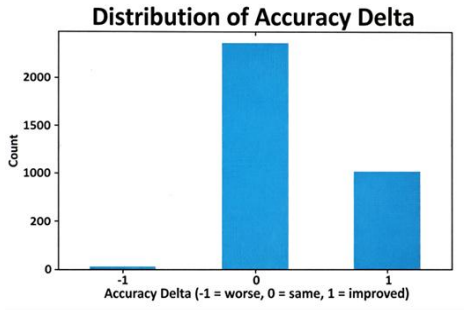

Fig.5. Accuracy-delta distribution

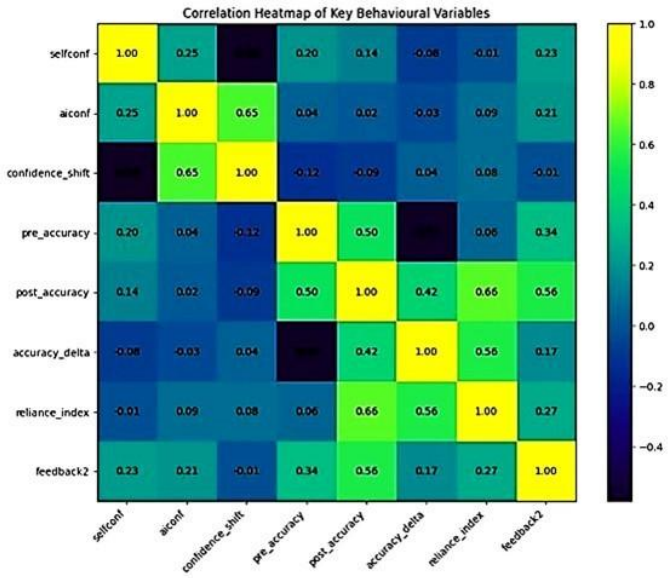

Fig.6. Correlation Heatmap

4.5 HYPOTHESIS TESTING RESULTS

- **Hypothesis 1**: AI Feedback Increases Decision Accuracy

This hypothesis is fully supported. A paired-samples t-test made it evident that accuracy increased significantly from 0.44 before feedback to 0.72 after feedback, with t = −35.53, p < 0.001, and Cohen's d = 0.61. A medium-to-large effect is indicated by the result, and confirmation is provided that decision quality in the task environment was raised to a higher level by AI feedback. Direct support for the first objective, assessing whether AI feedback leads to better performance is provided by this finding.

- **Hypothesis 2**: Misleading AI Feedback Reduces Confidence

Only partial support for this hypothesis is provided by the evidence. Post-decision confidence was significantly brought down to a lower level under misleading feedback (M = 0.49) than under accurate feedback (M = 0.59), with t = -12.34, p = 0.001, and d = -0.42. However, no meaningful difference between the two conditions was found in the average confidence shift (t = 0.51, p = 0.61, d = 0.02). Therefore, a lower final confidence level was produced by misleading feedback, but no significant change in the size of the confidence adjustment within a trial was produced by it. An important subtle but important distinction is provided by this result because an influence of feedback quality on confidence is seen to be present, but not as strongly or as consistently as accuracy.

Table.3. Independent-Samples T-Test for Confidence by Feedback Type

Outcome	Accurate AI Mean	Misleading AI Mean	t	p-value	Cohen's d
Post-decision confidence	0.59	0.49	-12.34	0.001	-0.42
Confidence Shift	0.06	0.07	0.51	0.61	0.02

- **Hypothesis 3**: Higher AI Reliance Reduces Independent Accuracy

This hypothesis was not supported in the expected negative direction. Across all three accuracy definitions, reliance on AI was associated with better performance rather than weaker performance. Under the broad accuracy measure, participants who followed the AI were correct 94.5% of the time, compared with 40.6% for non-followers, with OR = 25.12 and r = 0.58. Under the positive-score metric, the corresponding values were 57% versus 30%, with OR = 3.03 and r = 0.27. Under the strict best-move definition, AI followers were correct 57% of the time compared with 15% for non-followers, with OR = 7.70 and r = 0.44. The evidence therefore suggests that reliance on accurate AI advice improves rather than harms decision quality.

Table.4. Summary of Accuracy Outcomes for Reliance-Based Hypothesis Testing

Accuracy definition	AI followers	Non-followers	Correlation	Odds ratio
Broad accuracy	94.5%	40.6%	0.58	25.12
Positive-score accuracy	57%	30%	0.27	3.03
Best-move accuracy	57%	15%	0.44	7.70

4.6 PREDICTIVE MODELLING RESULTS

The machine-learning analysis was used to assess whether confidence, accuracy, and trust behaviour could be predicted from the available behavioural variables. For confidence shift, the regression models achieved extremely low prediction error, with RMSE values close to zero. This is unsurprising because confidence shift is mechanically defined by pre- and post-confidence. The Random Forest model assigned almost all predictive importance to selfconf and aiconf, while the remaining variables contributed negligibly.

Exceptional performance, with F1-scores above 0.99, was achieved by the classification models for decision accuracy. The Random Forest feature importance analysis made it clear that the principal predictors were reliance index (˜45%), pre accuracy (˜26%), feedback2 (˜24%), and AI suggestion ranking (˜2%). Less than 2% of total predictive power was added by confidence related variables. Central support for the paper's argument is provided by this result: objective accuracy is driven primarily by

trust behaviour and feedback correctness rather than by subjective confidence.

Table.5. Machine-Learning Performance Summary

Outcome predicted	Model	Metric	Score
Decision accuracy	Random Forest	F1-score	0.992
Decision accuracy	Gradient Boosting	F1-score	0.993
Decision accuracy	Logistic Regression	F1-score	0.993
Trust behaviour (reliance)	Random Forest	Accuracy	0.92
Trust behaviour (reliance)	Random Forest	F1-score	0.93

Strong performance, with accuracy scores between 0.92 and 0.93 and similarly high F1-scores, was also achieved by the reliance models. The most influential predictors were identified as post accuracy (˜32%), accuracy delta (˜22%), feedback2 (˜15%), and AI suggestion ranking (˜15%). Only a minor role was taken on by confidence variables, with less than 3% being accounted for by them. Reinforcement for the conclusion that trust behaviour is determined primarily by realised performance gains and feedback reliability is provided by this.

4.7 CLUSTERING AND BEHAVIOURAL PROFILES

To identify heterogeneity in human-AI interaction, the analysis used K-Means and DBSCAN clustering on confidence, accuracy, and reliance features. The K-Means results grouped participants into three broad behavioural profiles. These profiles are useful because they show that users do not respond uniformly to AI feedback; instead, they differ in how they calibrate trust and incorporate recommendations.

Table.6. K-Means Cluster Profiles

Cluster	Behavioural profile	Reliance level	Accuracy delta	Confidence shift
0	Over-reliant but improved users	High	Positive	Moderate
1	Resistant and low-performing users	Low	Negative	High
2	Adaptive high-performing users	Moderate	Stable	Low

Participants who depended to a great extent on the AI and achieved strong performance gains were captured by Cluster 0. Users who were strongly opposed to AI advice and tended to perform poorly were used as a depiction of Cluster 1. Adaptive users who appeared to use AI selectively and kept performance at a high level with relatively stable confidence were used as a depiction of Cluster 2. Support for the paper's objective of identifying distinct cognitive adaptation profiles is provided by the existence of these groups.

Smaller micro-patterns that were not visible in aggregate statistics were further revealed by DBSCAN. Hypertrusting users, hyper-resistant users, calibrated expert users, and volatile-confidence users were identified as the main behavioural types. Emphasis on a second important result is provided by these patterns: heterogeneous user responses to AI feedback are indicated, and important behavioural differences can be hidden from view by aggregate averages. Four behavioural microclusters were identified by the DBSCAN analysis. AI recommendations were consistently followed by hyper-trusting users, and high accuracy was achieved by them when the AI was correct. AI guidance tended to be ignored by hyper-resistant users, and persistently low accuracy was observed over time in this group. High accuracy both before and after feedback was made evident by calibrated expert users, and selective AI use was practiced by them. Finally, large fluctuations in confidence that were unrelated to actual performance were marked by volatile confidence users.

4.8 DISCUSSION OF KEY FINDINGS

Stronger decision performance under AI feedback, especially when calibrated reliance rather than blind trust or categorical rejection is used as the response, is demonstrated by the evidence taken together. The strongest and most consistent effects were found in the results for accuracy and reliance behaviour, while only modest change was seen in confidence. The conclusion that subjective confidence is not the main mechanism through which AI feedback improves performance is supported by this pattern. Instead, feedback correctness, reliance behaviour, and accuracy gains achieved in practice are identified as the decisive factors.

Practical implications for AI decision-support systems are suggested by this pattern. Reliable, performance-based trust calibration rather than simple increases in user confidence should be encouraged through system design. Reinforcement for this conclusion is provided by the clustering results through the demonstration that distinct behavioural classes including adaptive, resistant, and over reliant groups are occupied by users. Optimal performance is therefore unlikely to be achieved by a one-size-fits-all AI interface. Instead, support for calibrated reliance, adaptation to user behaviour, and preservation of human oversight in high-stakes decision environments should be built into system design.

5. CONCLUSION

How AI-generated feedback influences human confidence, decision accuracy, and reliance behaviour in a chess-based decision-making task was investigated in this study. Substantial improvement in decision accuracy under AI feedback, particularly when the feedback is correct, is demonstrated by the results, while only modest changes in confidence are seen by comparison. A closer link between trust behaviour and performance gains seen in the data as well as feedback correctness than with subjective confidence is also established by the analysis. Meaningful behavioural heterogeneity was brought to light by the clustering results, with adaptive, resistant, and over-reliant user profiles being identified. A contribution to the human–AI interaction literature is made by these findings through the demonstration that performance based trust calibration is more important than confidence alone for explaining effective decision support. A threefold structure is given to the study's main contribution. First, trial-level evidence that decision quality in complex problem-solving settings can be improved by accurate AI feedback is provided. Second, the finding that objective performance is predicted weakly by confidence relative to

reliance and feedback correctness is made clear by the evidence. Third, the ability of machine learning and clustering methods to bring to the surface behavioural patterns that are not visible through aggregate statistics alone is demonstrated, with a more detailed account of how users interact with AI systems being provided as a result. Several limitations are acknowledged by the study. Reliance on secondary data from a chess-based task is built into the study design, and generalizability to other domains such as healthcare, finance, or engineering may therefore be limited.

Extension of this work in several directions is recommended for future research. Examination of how trust calibration evolves across repeated exposure to AI feedback could be carried out through longitudinal studies. Testing of whether the same confidence accuracy reliance relationships hold outside chess-based decision tasks could also be performed with domain-specific datasets from higher-stakes environments. Incorporation of richer behavioural and psychological variables, such as experience level, task difficulty, and cognitive workload, may be pursued in future work to improve prediction and interpretation. In addition, refinement of feedback design so that calibrated reliance is supported while human autonomy and oversight are preserved could be enabled through the use of interpretable AI methods.

REFERENCES

[1] B. Sahoh and A. Choksuriwong, "The Role of Explainable Artificial Intelligence in High-Stakes Decision-Making Systems: A Systematic Review", *Journal of Ambient Intelligence and Humanized Computing*, Vol. 14, No. 6, pp. 7827-7843, 2023.

[2] M. Khosravi, Z. Zare and R. Izadi, "Artificial Intelligence and Decision-Making in Healthcare: A Thematic Analysis of a Systematic Review of Reviews", *Health Services Research and Managerial Epidemiology*, Vol. 11, pp. 1-24, 2024.

[3] Y. Lu, "The Current Status and Developing Trends of Industry 4.0: A Review", *Information Systems Frontiers*, Vol. 27, No. 1, pp. 215-234, 2025.

[4] M. Alabi, "*Adoption of Artificial Intelligence in Business: Challenges and Strategic Implementation*", Springer, 2025.

[5] M. Steyvers and A. Kumar, "Three Challenges for AI-Assisted Decision-Making", *Perspectives on Psychological Science*, Vol. 19, No. 5, pp. 722734-722745, 2024.

[6] H. Tejeda Lemus, A. Kumar and M. Steyvers, "*How Displaying AI Confidence Affects Reliance and Hybrid Human-AI Performance*", IOS Press, 2023.

[7] N. Ishizu, W.L. Yeoh, H. Okumura and O. Fukuda, "The Effect of Communicating AI Confidence on Human Decision Making when Performing a Binary Decision Task", *Applied Sciences*, Vol. 14, No. 16, pp. 7192-7205, 2024.

[8] G. Romeo and D. Conti, "Exploring Automation Bias in Human-AI Collaboration: A Review and Implications for Explainable AI", *AI and SOCIETY*, Vol. 41, pp. 259-278, 2025.

[9] J. Han and D. Ko, "Trust Formation, Error Impact, and Repair in Human-AI Financial Advisory: A Dynamic Behavioral Analysis", *Behavioral Sciences*, Vol. 15, No. 10, pp. 1370-1387, 2025.

[10] A. Rechkemmer and M. Yin, "When Confidence Meets Accuracy: Exploring the Effects of Multiple Performance Indicators on Trust in Machine Learning Models", *Proceedings of International Conference on Human Factors in Computing Systems*, pp. 1-14, 2022.

[11] Y. Zhang, Q.V. Liao and R.K. Bellamy, "Effect of Confidence and Explanation on Accuracy and Trust Calibration in AI Assisted Decision Making", *Proceedings of International Conference on Fairness, Accountability, and Transparency*, pp. 295-305, 2020.

[12] R.A. McKinnon and R.D. Taylor, "Automation Bias and Decision-Making Performance", *Human-Computer Interaction*, Vol. 36, No. 1, pp. 4559-4568, 2021.

[13] M. Vaccaro, A. Almaatouq and T. Malone, "When Combinations of Humans and AI are Useful: A Systematic Review and Meta-Analysis", *Nature Human Behaviour*, Vol. 8, No. 12, pp. 2293-2303, 2024.

[14] H. Ruschemeier, "*Generative AI and Data Protection in Cambridge Forum on AI: Law and Governance*", Cambridge University Press, 2025.

[15] H. Li, E.R. Musaev and C. Zhang, "Artificial Intelligence in Surgery: Evolution, Trends, and Future Directions", *International Journal of Surgery*, Vol. 111, No. 2, pp. 2101-2111, 2025.

[16] . Sakamoto, Y. Harada and T. Shimizu, "Facilitating Trust Calibration in Artificial Intelligence-–Driven Diagnostic Decision Support Systems for Determining Physicians' Diagnostic Accuracy: Quasi Experimental Study", *JMIR Formative Research*, Vol. 8, No. 1, pp. 58666-58675, 2024.

[17] D. Ma, H. Akram and I.H. Chen, "Artificial Intelligence in Higher Education: A Cross-Cultural Examination of Students' Behavioural Intentions and Attitudes", *International Review of Research in Open and Distributed Learning*, Vol. 25, No. 3, pp. 134-157, 2024.

[18] D.O.T. Oyekunle, U. Okwudili Matthew and D. Boohene, "Trust Beyond Technology Algorithms: A Theoretical Exploration of Consumer Trust and Behavior in Technological Consumption and AI Projects", *Journal of Computer and Communications*, Vol. 12, No. 6, pp. 4210-4236, 2024.

[19] J. Li, Y. Yang, Q.V. Liao and Y.C. Lee, "As Confidence Aligns: Understanding the Effect of AI Confidence on Human Self-confidence in Human-AI Decision Making", *Proceedings of CHI Conference on Human Factors in Computing Systems*, pp. 1-16, 2025.

[20] A. Sakamoto, W. Fujita, Y. Nakamura, N. Yokotsuka and N. Kagiyama, "Artificial Intelligence in Echocardiography: Current Applications and Future Perspectives", *Journal of Echocardiography*, Vol. 35, pp. 1-10, 2025.

[21] S.A. Khanday, K.M. Vali, M. Junaid and M. Ahmad, "Understanding Positivism: A Qualitative Exploration of Its Principles and Relevance Today", *European Journal of Applied Sciences*, Vol. 12, No. 6, pp. 1-14, 2024.

[22] L. Chong, G. Zhang and J. Cagan, "Data on Human Decision, Feedback, and Confidence during an Artificial Intelligence-Assisted Decision-Making Task", *Data in Brief*, Vol. 46, No. 1, pp. 108884-108895, 2023.