Applied Artificial Intelligence Research

INNOVIAIR

AI Transparency and Employee Innovation: The Mediating Role of Psychological Safety and the Moderating Effect of AI Self-Efficacy

Qian Li[a,*], Peilin Li[a], Chan Sai Keong[a]

[a]*Faculty of Business, Information & Human Sciences, Kuala Lumpur University of Science and Technology, Kajang, Selangor 43000, Malaysia*

ARTICLEINFO

Keywords:
AI Transparency
Psychological Safety
Employee Innovation
AI Self-Efficacy

ABSTRACT

As AI technologies become more integrated into everyday organizational workflows, it is essential to understand how employees interpret and interact with these systems to support innovative outcomes.This study investigates how transparency in AI systems influences employee innovation, considering the role of psychological safety and individual confidence in using AI. Data were gathered from 447 knowledge workers in small and medium-sized enterprises in China who regularly interact with AI tools. Analysis using structural equation modeling shows that clearer AI processes enhance employees' sense of psychological safety, which in turn supports innovative actions. Moreover, employees with greater confidence in using AI benefit more from transparency, amplifying its positive impact on innovation. The results underscore that promoting innovation requires not only transparent system design but also initiatives that strengthen employees' AI skills and create psychologically supportive environments. This research contributes to understanding the human side of AI adoption by linking system transparency to practical organizational outcomes and offering guidance for effective AI–human integration in workplaces.

1.Introduction

Artificial intelligence (AI) is increasingly reshaping contemporary workplaces, altering how tasks are performed, how decisions are made, and how employees engage with organizational processes. As organizations adopt AI technologies to enhance efficiency, flexibility, and innovation, employees are increasingly required to interact with AI systems in their daily roles[1]. However, AI adoption is not purely a technical matter; it also introduces new psychological and organizational challenges. Employees often face uncertainty regarding how AI systems function, whether outputs are reliable, and how AI-mediated processes affect their experiences and well-being[2]. Empirical evidence indicates that despite widespread AI deployment, many employees do not fully understand system operations, which can shape their evaluations and responses to AI in workplace contexts[3]. Consequently, effective AI implementation depends not only on system integration but also on employees' perceptions, interpretations, and experiences of AI-enabled work.

Research examining AI's impact on employee attitudes, behaviors, and psychological experiences has emphasized the role of AI transparency. Transparency refers to the extent to which users can comprehend the logic, processes, and rationale behind AI outputs[4,5]. Studies indicate that when employees better understand how algorithmic decisions are generated, they tend to demonstrate greater trust, acceptance, and confidence in AI systems, though these effects may vary across organizational and cultural contexts[6]. This focus aligns with the broader discourse in Applied Artificial Intelligence Research, which underscores the practical importance of explainable AI, model interpretability, and the downstream behavioral consequences of AI applications in areas such as recommendation systems, credit scoring, and operational decision-making[7-9].

In addition to shaping trust and system acceptance, AI-driven work environments can influence employees' psychological states—such as their sense of security, autonomy, and engagement—by modifying how tasks are structured, how performance is monitored, and how workflows are organized. Research in organizational behavior has consistently emphasized that psychological safety—the

* Corresponding author.
E-mail address: 549863920@qq.com.
https://doi.org/10.65455/dxhkxd06
Received 25 March 2026; Received in revised form 11 April 2026; Accepted 16 April 2026; Available 19 May 2026
© The Author(s) 2026. This article is published with open access in Applied Artificial Intelligence Research under a Creative Commons Attribution 4.0 International License (CC BY 4.0). https://creativecommons.org/licenses/by/4.0/

conviction that one can take interpersonal risks without facing negative repercussions—is fundamental for encouraging initiative, experimentation, learning, and innovative behavior. Furthermore, employees' beliefs in their own efficacy determine how effectively they can convert workplace conditions into tangible innovative outcomes. Nevertheless, existing studies have yet to thoroughly examine the specific conditions or individual characteristics under which AI transparency yields beneficial psychological and behavioral effects.

To address these gaps, this study develops and empirically tests a moderated mediation model linking AI transparency, psychological safety, AI self-efficacy, and employee innovation. Specifically, the research investigates whether perceived AI transparency enhances employee innovation, whether psychological safety mediates this relationship, and whether AI self-efficacy strengthens the positive effect of transparency on psychological safety. Knowledge workers in Chinese small and medium-sized enterprises (SMEs), who frequently interact with AI tools, were selected as the target population to provide a realistic context for examining the translation of system transparency into innovative behaviors via psychological mechanisms and capability beliefs.

This study contributes in three primary ways. First, it extends the literature on AI transparency beyond trust- and acceptance-related outcomes to explore its influence on employee innovation. By focusing on behavioral consequences, the study broadens understanding of the organizational impact of transparent AI systems. Second, it identifies psychological safety as a key mechanism linking transparency to innovation, integrating research on AI-enabled work with established organizational behavior theory. Third, it introduces AI self-efficacy as a boundary condition, demonstrating that the benefits of transparency depend on employees' confidence in interpreting and applying AI outputs. In doing so, the study provides a more nuanced, employee-centered perspective on human–AI interaction in workplace settings.

Practically, the findings highlight that fostering innovation through AI requires more than simply deploying transparent systems. Organizations should combine system design with initiatives that enhance employee capabilities and create psychologically safe environments, ensuring that employees can interpret, question, and apply AI-generated insights effectively.

2.Literature review and hypotheses

2.1.AI transparency and psychological safety

AI transparency describes how clearly employees can grasp the reasoning, procedures, and results produced by AI systems. In modern workplaces, this clarity is increasingly important, as workers often depend on AI tools to guide decisions, complete tasks, and assess performance-related information. When employees have a clear understanding of how AI arrives at its outputs, they can evaluate these results more effectively and use them proactively, turning AI from an opaque mechanism into a practical aid for informed decision-making[10,11].

Research in explainable AI emphasizes that transparency improves user understanding, reduces ambiguity, and fosters trust, aligning with broader technology acceptance frameworks that highlight perceived usefulness and comprehension as critical antecedents of user responses[12]. Applied AI studies reinforce this perspective, showing that interpretability enhances outcomes in domains such as telecom churn prediction, credit scoring, and algorithmic recommendation systems[7-9]. In practice, the effectiveness of AI increasingly depends not only on predictive accuracy but also on employees' capacity to interpret outputs and translate them into actionable knowledge.

From a psychological perspective, unclear or opaque AI processes generate uncertainty, which can reduce confidence and increase hesitation when interacting with systems. Transparent AI, by contrast, clarifies operational logic, enabling employees to anticipate how outputs are generated and how they should be interpreted. This predictability reduces perceived risk, lowers cognitive load, and fosters a sense of safety when engaging with AI[13,14].

Psychological safety, defined as the perception that one can take interpersonal and task-related risks without fear of negative consequences, is particularly relevant in AI-enabled workplaces. When AI systems are opaque, employees may feel constrained in questioning recommendations or verifying outputs. Transparent AI, however, empowers employees to interpret, discuss, and critically engage with system outputs, fostering a climate of safety and support. Consequently, AI transparency is likely to enhance psychological safety by reducing dependency on inaccessible systems and legitimizing questioning behaviors.

H1: AI transparency is positively associated with psychological safety.

2.2.Psychological safety and employee innovation

Employee innovation refers to the process of generating, promoting, and implementing new and useful ideas within one's job responsibilities. Beyond simply being creative, innovative behavior also involves actively turning ideas into practical solutions[15]. Because innovation often entails uncertainty and potential social or professional consequences, employees are more likely to take initiative when they believe that expressing ideas or experimenting will not lead to negative judgment or personal repercussions.

A sense of psychological safety plays a central role in enabling such innovative behavior. Work environments that support open communication, constructive challenge, and experimentation allow employees to explore novel approaches without fear of criticism or sanction[16]. Individuals who perceive a secure and supportive climate are more inclined to offer unconventional solutions, test new methods, and share knowledge that can enhance work processes[17].

In AI-enabled workplaces, psychological safety may play an even more critical role. Employees who understand AI systems and feel supported in questioning outputs can transform system transparency into concrete innovative behaviors. By mitigating concerns about mistakes, criticism,

or social judgment, psychological safety allows employees to experiment, learn, and apply novel ideas with greater confidence.

H2: Psychological safety is positively associated with employee innovation.

2.3.Direct influence of AI transparency on innovation

In addition to its indirect effect through psychological safety, AI transparency may also directly facilitate innovative behavior. When employees understand the logic behind AI outputs, they are better equipped to integrate machine-generated insights with domain knowledge. This enables critical evaluation, adaptive problem-solving, and the implementation of creative solutions. In essence, transparency reduces cognitive barriers to effective human–AI collaboration and creates conditions conducive to proactive and reflective work behavior[18,19,20].

Empirical studies in algorithmic management and explainable AI support the notion that interpretability enhances employee proactivity, allowing workers to question, refine, and apply system outputs in innovative ways[11,21,22]. Employees who understand how AI generates suggestions are more likely to leverage these outputs creatively, making transparency a direct contributor to innovation in addition to its indirect effect via psychological safety.

H3: AI transparency is positively associated with employee innovation.

2.4.Mediating role of psychological safety

Although transparency may influence innovation directly, the mechanism by which this occurs is critical to understand. Technological features of the workplace often shape behavior through proximal psychological processes rather than exerting purely direct effects. In this study, AI transparency is conceptualized as a feature of the work environment, while psychological safety represents a cognitive–emotional state that enables employees to act innovatively.

Transparent AI systems enhance predictability and understandability, reducing uncertainty and supporting a sense of control over work outcomes. Psychological safety, in turn, encourages employees to speak up, take initiative, and experiment without fear of interpersonal consequences, thereby translating technological clarity into innovative behavior[23,24].

H4: Psychological safety mediates the relationship between AI transparency and employee innovation.

2.5.Moderating role of AI self-efficacy

Employees vary in how effectively they can make use of information provided by AI transparency, depending on their confidence in using AI tools. According to social cognitive theory, individuals' beliefs about their capabilities shape how they process information, persist in demanding tasks, and engage with complex challenges[25]. Those with higher confidence in applying AI are better equipped to interpret transparency cues, manage uncertainty, and interact with AI systems proactively, thereby strengthening the positive impact of transparency on their sense of psychological safety.

Conversely, employees with low self-efficacy may struggle to benefit from transparent systems, perceiving information as confusing or insufficiently actionable, which attenuates its positive impact on psychological safety[26,27]. Therefore, AI self-efficacy acts as a boundary condition shaping the effectiveness of transparency interventions in the workplace.

H5: AI self-efficacy positively moderates the relationship between AI transparency and psychological safety, such that the effect is stronger when AI self-efficacy is high.

2.6.Moderated mediation hypothesis

If psychological safety mediates the transparency–innovation relationship, and AI self-efficacy strengthens the transparency → safety path, the indirect effect of transparency on innovation should vary according to levels of self-efficacy. Employees with high AI self-efficacy are expected to translate transparency into a greater sense of safety, which then fosters stronger innovative behavior. In contrast, low self-efficacy may weaken this pathway.

H6: The indirect effect of AI transparency on employee innovation through psychological safety is stronger at higher levels of AI self-efficacy.

2.7.Conceptual model

Figure 1 illustrates the hypothesized relationships among AI transparency, psychological safety, AI self-efficacy, and employee innovation. Transparency is proposed to influence innovation both directly and indirectly through psychological safety, while AI self-efficacy moderates the transparency → safety link, shaping the overall mediated effect. This model integrates cognitive, psychological, and capability-based perspectives, highlighting the contingent nature of technological effects on employee innovation[28].

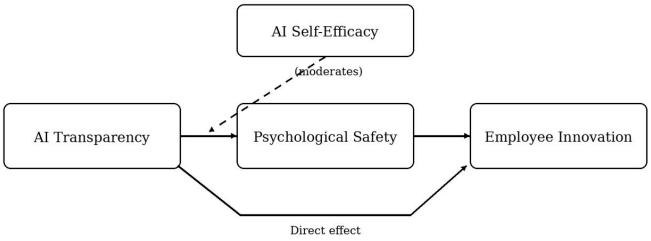


Fig 1. Conceptual model (AI self-efficacy moderates the relationship between AI transparency and psychological safety)

3.Methodology

3.1.Research design and sample

To examine the proposed relationships among AI transparency, psychological safety, AI self-efficacy, and employee innovation, this study employed a quantitative survey methodology. The target population comprised full-time knowledge workers in Chinese small and medium-sized enterprises (SMEs) who regularly interacted with AI tools as part of their daily responsibilities. Knowledge-intensive roles

were chosen because these employees frequently rely on AI systems for decision-making, data analysis, and workflow management, providing an appropriate context for studying innovation-related behaviors influenced by AI.

Data collection occurred between March and June 2025 via an online questionnaire distributed through professional networks and organizational channels. Participation criteria required respondents to report at least weekly interaction with AI-based tools over the previous three months. This approach ensured that the collected responses reflected actual experiences rather than hypothetical perceptions or limited exposure.

Of the 500 questionnaires distributed, 447 valid responses were retained after excluding incomplete surveys and participants failing to meet the screening criteria, resulting in an effective response rate of 89.4%. The sample size exceeded recommended thresholds for structural equation modeling, offering sufficient statistical power to examine direct, mediating, moderating, and conditional indirect effects. Respondents were informed that participation was voluntary, anonymous, and exclusively for academic purposes, following ethical guidelines and minimizing social desirability bias.

3.2.Measurement instruments

All constructs were measured using established multi-item scales adapted to the AI-enabled workplace context. Items were slightly reworded for relevance while preserving conceptual fidelity. Unless otherwise noted, all responses were recorded on five-point Likert scales ranging from 1 ("strongly disagree") to 5 ("strongly agree").

AI Transparency was assessed with five items adapted from prior explainable AI research, capturing employees' understanding of system outputs and decision logic[4,7]. A representative item reads, "I understand how the AI system generates its recommendations." Higher scores indicate greater perceived transparency.

Psychological Safety was measured with a five-item scale based on Edmondson[29], reflecting employees' perceptions of whether it is safe to share ideas, question procedures, and take interpersonal risks. A sample item is, "I feel comfortable raising concerns about work processes without fear of negative consequences."

Employee Innovation was evaluated using five items from Scott and Bruce[30], capturing the frequency of generating, promoting, and implementing novel ideas. A representative item is, "I frequently propose creative solutions to work challenges."

AI Self-Efficacy was measured using five items adapted from Compeau and Higgins[26], reworded for AI-related tasks. The scale assesses confidence in effectively using AI tools. For example, "I am confident in my ability to use AI tools to complete work tasks efficiently."

3.3.Control variables

To minimize potential confounding influences, demographic factors such as gender, age, education level, and tenure within the organization were incorporated into the analyses. These variables have been commonly used in previous studies examining technology adoption, innovation, and employee behaviors[31].

3.4.Common method bias control

Since all constructs were self-reported in a single survey, several procedural and statistical measures were implemented to mitigate common method bias. Respondents completed the survey anonymously, item sequences were randomized, and instructions clarified that there were no correct or incorrect answers. Statistically, variance inflation factors (VIFs) were examined to assess multicollinearity, with all values below 3.3[32], suggesting that collinearity and method bias were unlikely to distort the results.

3.5.Data analysis

The dataset was processed using SPSS 26 and SmartPLS 4. Partial least squares structural equation modeling (PLS-SEM) was chosen because it effectively handles complex models with multiple latent variables, direct and indirect effects, and moderation or conditional pathways. This method is also resilient to deviations from normality, making it suitable for exploratory and predictive analyses of technology-related behaviors[33,34].

Preliminary Examination: Descriptive statistics and correlation matrices were calculated to inspect the distributions, detect potential outliers, and gain an initial understanding of the relationships among constructs.

Assessment of the Measurement Model: Reliability was evaluated using Cronbach's alpha and composite reliability, with values above 0.70 considered satisfactory. Convergent validity was assessed through the average variance extracted, with a benchmark of 0.50, while discriminant validity was examined using the heterotrait–monotrait ratio to ensure constructs were empirically distinct[35].

Evaluation of the Structural Model: Hypothesized links between constructs were tested via bootstrapping with 5,000 resamples. Coefficients, t-values, p-values, and confidence intervals were reviewed to determine significance. Model fit and explanatory power were quantified using R^2, and predictive relevance was checked with the Stone–Geisser Q^2 statistic. Collinearity diagnostics were also performed to rule out bias due to multicollinearity.

Mediation Analysis: The indirect influence of AI transparency on employee innovation through psychological safety was estimated using bootstrapped confidence intervals. Mediation effects were considered statistically meaningful if the 95% confidence interval excluded zero.

Moderation Analysis: AI self-efficacy was included as a moderator by generating an interaction term with AI transparency. A significant coefficient indicates that the strength of the transparency → psychological safety relationship varies across efficacy levels.

Moderated Mediation Analysis: Conditional indirect effects were examined to assess whether the mediated effect of AI transparency on innovation through psychological safety varies across levels of AI self-efficacy, following established procedures for testing moderated mediation models[36,37].

3.6.Methodological justification

The selected methodology aligns with the study's objectives by capturing realistic AI interactions in knowledge-intensive SMEs and leveraging PLS-SEM to accommodate latent constructs, interaction effects, and complex indirect pathways. Measurement instruments were drawn from validated scales and adapted for contextual relevance. Procedural safeguards, control variables, and rigorous statistical evaluation collectively enhance the validity, reliability, and replicability of the findings.

4.Results

4.1.Descriptive statistics and preliminary correlations

Table 1 presents descriptive statistics and zero-order correlations for the four primary constructs. All measures employed five-point Likert scales. Mean values ranged from 2.97 to 3.11, with standard deviations between 0.89 and 0.91, indicating moderate response dispersion and avoiding ceiling or floor effects.

Correlation analyses offered initial support for the theoretical model. AI transparency correlated positively with psychological safety (r = 0.389) and employee innovation (r = 0.333). Psychological safety demonstrated the strongest bivariate association with innovation (r = 0.516), consistent with its hypothesized mediating function. In contrast, AI self-efficacy displayed a weak zero-order correlation with innovation (r = 0.071), implying that its influence is contingent rather than directly predictive. Because PLS-SEM does not require multivariate normality, no strict assumptions about data distribution were imposed.

Table 1. Descriptive statistics and correlation matrix

Construct	Mean	SD	1	2	3	4
AI Transparency	3.02	0.89	1.000			
Psychological Safety	3.11	0.90	0.389	1.000		
AI Self-Efficacy	2.97	0.89	0.312	0.207	1.000	
Innovation	3.08	0.91	0.333	0.516	0.071	1.000

4.2.Measurement model evaluation

Before testing the hypotheses, the measurement model was assessed to ensure that the latent constructs were both reliable and valid. This evaluation considered factor loadings, the consistency of the items, the extent to which constructs captured their intended variance, and the distinctiveness between constructs. Overall, the findings indicated that the measures were psychometrically sound and suitable for further structural analysis.

4.2.1.Indicator loadings, reliability, and convergent validity

The alignment of each survey item with its corresponding construct was evaluated, and all items showed strong associations, with standardized loadings exceeding 0.70. Reliability of the constructs was examined using both Cronbach's alpha and composite reliability (CR). As reported in Table 2, alpha coefficients ranged from 0.866 to 0.887, while CR values were between 0.903 and 0.917, indicating that the items consistently measure their intended constructs.

Convergent validity was also assessed by calculating the average variance extracted (AVE), which fell between 0.652 and 0.690, surpassing the standard 0.50 benchmark. These results collectively indicate that the measurement items provide a valid and coherent representation of their constructs, confirming that the model is suitable for further structural analysis.

Table 2. Measurement model reliability and convergent validity

Construct	Items	Cronbach's α	CR	AVE
AI Transparency	5	0.866	0.903	0.652
Psychological Safety	5	0.880	0.912	0.675
AI Self-Efficacy	5	0.869	0.905	0.656
Innovation	5	0.887	0.917	0.690

Convergent validity was assessed through average variance extracted (AVE). The AVE values ranged from 0.652 to 0.690, all of which exceeded the recommended cutoff of 0.50. Taken together, these results indicate that the indicators adequately capture their intended latent constructs and that the measurement model demonstrates acceptable reliability and convergent validity.

4.2.2.Discriminant validity

To evaluate whether the constructs were distinct from one another, the heterotrait–monotrait (HTMT) ratio was calculated. All computed values fell well below the commonly accepted cutoff of 0.85, with the highest ratio reaching 0.584. These results indicate that each construct captures unique variance and that there is no significant overlap between them.

Table 3. Discriminant validity assessment (HTMT)

Construct	AT	PS	SE	IN
AI Transparency (AT)	1.000	0.446	0.360	0.380
Psychological Safety (PS)	0.446	1.000	0.237	0.584
AI Self-Efficacy (SE)	0.360	0.237	1.000	0.089
Innovation (IN)	0.380	0.584	0.089	1.000

4.3.Structural model assessment

Once the measurement model was confirmed to be reliable and valid, attention shifted to the structural model. The goal was to examine the relationships among predictors, evaluate how well the model accounts for variation in outcomes, assess its ability to make accurate predictions, and determine the strength of estimated effects.

4.3.1.Collinearity diagnostics

Table 4. Collinearity assessment (VIF)

Predictor	VIF
AT	2.12
SE	1.18
AT×SE	2.41
PS	1.76

Variance inflation factors (VIFs) were calculated to identify whether predictor variables might be excessively correlated. Observed values fell between 1.18 and 2.41, well below the conventional 3.3 benchmark. This indicates that

multicollinearity is unlikely to distort the estimated paths, and any bias from common method variance is minimal.

4.3.2.Explanatory and predictive capacity

The model explained 23.1% of the variance in psychological safety and 30.2% in innovation (R^2), indicating meaningful explanatory power for organizational behavior research. Predictive relevance, assessed via Stone–Geisser Q^2, was positive for both constructs (psychological safety Q^2 = 0.214; innovation Q^2 = 0.269), suggesting reliable out-of-sample predictive potential.

Table 5. Explanatory power (R^2) and predictive relevance (Q^2)

Endogenous construct	R^2	Q^2
Psychological Safety	0.231	0.214
Innovation	0.302	0.269

4.3.3.Direct effects and path coefficients

To evaluate the hypothesized relationships, a bootstrapping procedure with 5,000 resamples was conducted. Results showed that AI transparency had a significant positive impact on psychological safety ($\beta = 0.352$, $p < 0.001$), confirming H1. In turn, psychological safety exerted a positive influence on employee innovation ($\beta = 0.452$, $p < 0.001$), supporting H2. AI transparency also maintained a direct positive association with innovation ($\beta = 0.167$, $p < 0.001$), in line with H3.

Additionally, AI self-efficacy had a modest yet statistically significant positive effect on psychological safety ($\beta = 0.117$, $p = 0.015$). Notably, the interaction between AI transparency and AI self-efficacy significantly predicted psychological safety ($\beta = 0.295$, $p < 0.001$), indicating that the benefits of transparency are enhanced when employees possess higher confidence in using AI, consistent with H5.

Table 6. Structural model path coefficients

Hypothesis	Path	β	t	p	95% CI
H1	AT → PS	0.352	7.779	<0.001	[0.263, 0.441]
Control	SE → PS	0.117	2.441	0.015	[0.023, 0.211]
H5	AT×SE → PS	0.295	7.109	<0.001	[0.214, 0.376]
H2	PS → IN	0.452	10.282	<0.001	[0.365, 0.538]
H3	AT → IN	0.167	3.843	<0.001	[0.084, 0.253]

4.3.4.Effect sizes

Effect sizes (f^2) indicated that psychological safety had a medium effect on innovation (f^2 = 0.242). AI transparency exhibited small-to-medium effects on psychological safety (f^2 = 0.140) and a small effect on innovation (f^2 = 0.032). The interaction effect was small-to-medium (f^2 = 0.093), highlighting the central role of psychological safety and the moderating influence of AI self-efficacy.

Table 7. Effect size (f^2) assessment of structural paths

Path	f^2	Interpretation
AT → PS	0.14	small–medium
SE → PS	0.02	small
AT×SE → PS	0.09	small–medium
PS → IN	0.24	medium
AT → IN	0.03	small

Figure 2 shows the structural model results.

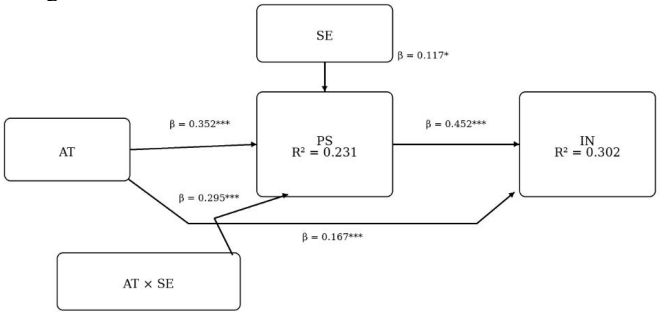


Fig 2. Structural model results

(Note: AT = AI transparency; SE = AI self-efficacy; PS = psychological safety; IN = innovation. Standardized coefficients are reported. *p < .05, **p < .01, ***p < .001.)

4.4.Mediation analysis

The mediating effect of psychological safety was tested via bootstrapped indirect estimates. Results showed a significant positive indirect effect of AI transparency on innovation through psychological safety ($\beta = 0.159$, 95% CI [0.110, 0.212]), supporting H4. The direct effect of AI transparency remained significant ($\beta = 0.167$, $p < 0.001$), indicating partial mediation.

Table 8. Mediation analysis results

Effect	Indirect path	Indirect β	95% CI	Result
H4	AT → PS → IN	0.159	[0.110, 0.212]	Supported

4.5.Moderation analysis

The analysis revealed that AI self-efficacy significantly influenced the strength of the relationship between AI transparency and psychological safety ($\beta = 0.295$, $p < 0.001$), consistent with H5. As illustrated in Figure 3, the effect of transparency on psychological safety becomes more pronounced as employees' confidence in using AI increases, with the slope steepening from low to high levels of self-efficacy.

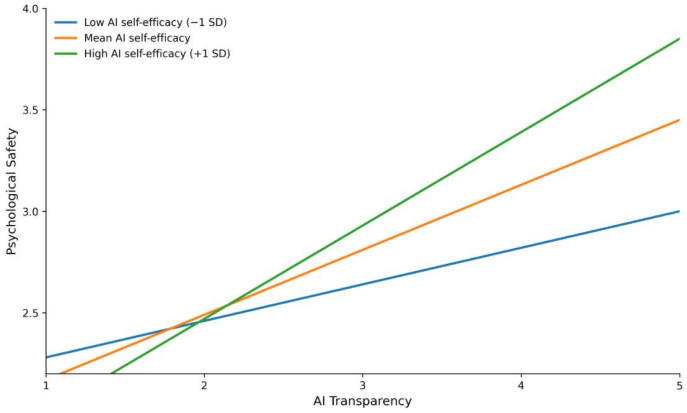


Fig 3. Simple slope plot of the interaction between AI transparency and AI self-efficacy on psychological safety

4.6.Moderated mediation analysis

Conditional indirect effects were computed at −1 SD, mean, and +1 SD levels of AI self-efficacy. The indirect effect

strengthened progressively with self-efficacy: weak and non-significant at low (β = 0.041), significant at mean (β = 0.159), and strongest at high levels (β = 0.278). These results confirm H6 and highlight that the effect of transparency on innovation via psychological safety is contingent on employee AI self-efficacy.

Table 9. Conditional indirect effects at different levels of AI self-efficacy

Level of AI self-efficacy	Conditional indirect effect (AT → PS → IN)	95% CI	Interpretation
Low (−1 SD)	0.041	[−0.014, 0.099]	weaker / marginal
Mean	0.159	[0.110, 0.212]	significant
High (+1 SD)	0.278	[0.209, 0.353]	stronger / significant

4.7.Summary of hypothesis testing

Table 10 summarizes results for all six hypotheses. Each was supported, indicating that AI transparency contributes to innovation directly and indirectly via psychological safety, with the mediated pathway amplified by AI self-efficacy. These findings provide comprehensive support for the proposed moderated mediation framework.

Table 10. Summary of hypothesis testing results

Hypothesis	Statement	Supported?
H1	AI transparency → psychological safety (+)	Supported
H2	Psychological safety → innovation (+)	Supported
H3	AI transparency → innovation (+)	Supported
H4	Psychological safety mediates AT → IN	Supported
H5	AI self-efficacy moderates AT → PS (+ stronger when SE high)	Supported
H6	Conditional indirect effect (moderated mediation)	Supported

5.Discussion

5.1.Overview of key findings

This study investigated a moderated mediation framework in which perceived AI transparency influences employee innovation, both directly and indirectly through psychological safety, while AI self-efficacy strengthens the transparency–safety relationship. Data collected from knowledge workers in Chinese SMEs were analyzed using PLS-SEM, confirming support for all hypothesized paths. Specifically, AI transparency positively predicted psychological safety, psychological safety significantly enhanced employee innovation, and transparency retained a direct effect on innovation after accounting for the mediator, indicating partial mediation.

Furthermore, AI self-efficacy amplified the positive effect of transparency on psychological safety. Consequently, the indirect influence of transparency on innovation via psychological safety was more pronounced when employees reported higher AI self-efficacy. These findings suggest that transparency enhances innovation not only by clarifying AI outputs but also by fostering an environment where employees feel secure to experiment, question, and engage creatively. Importantly, the effect of transparency is contingent on employees' confidence in working with AI, highlighting the interplay between technological design and individual capability beliefs.

5.2.AI transparency as an enabling work resource

The observed positive link between transparency and psychological safety suggests that AI transparency functions as an important work resource. Beyond technical performance, transparent AI systems provide interpretive clarity that reduces uncertainty and supports employees' comprehension of algorithmic outputs[4]. When outputs are understandable, employees are less likely to perceive AI as a "black box," reducing ambiguity in decision-making and improving perceived control over work outcomes.

From an organizational perspective, transparency should be understood as a cognitive and psychological resource rather than a purely technical feature. By enhancing predictability and interpretability of AI processes, transparency equips employees to engage safely with AI-assisted tasks, promoting proactive and innovative behavior. This aligns with broader evidence that employee experiences of AI-mediated work, rather than system functionality alone, drive the behavioral and psychological consequences of technology adoption[13].

5.3.Psychological safety as the core mechanism

Psychological safety was identified as the key mechanism through which AI transparency influences innovative outcomes. Among all the pathways in the model, it demonstrated the greatest impact on employees' innovative actions. When individuals perceive a work environment that supports open communication and experimentation, they are more willing to propose new ideas, question existing practices, and explore unconventional solutions without fearing negative consequences.

In AI-enabled work contexts, transparent systems enhance this sense of safety by allowing employees to interpret, question, and interact confidently with AI outputs. This finding bridges human–AI interaction research and organizational behavior theory, demonstrating that technology can shape innovation indirectly via psychological mechanisms. The partial mediation observed indicates that transparency also facilitates innovation directly, likely by reducing cognitive load, increasing interpretability, and supporting adaptive application of AI outputs. This dual pathway—psychological and cognitive—is consistent with research on explainable AI, which emphasizes that interpretability enhances both understanding and effective reliance on AI systems[18,19].

5.4.The moderating role of AI self-efficacy

The moderation analysis confirmed that AI self-efficacy strengthens the positive effect of transparency on psychological safety. Social cognitive theory posits that individuals' efficacy beliefs shape how they interpret information, respond to challenges, and persist in complex

tasks. In this study, transparency cues such as system explanations, rationales, and decision logic were most impactful when employees believed in their ability to comprehend and apply AI outputs.

Employees with low self-efficacy may find transparency cues difficult to interpret, limiting their effectiveness. This underscores that individual capability beliefs are not peripheral but critical in determining how technological features are experienced and leveraged for workplace behavior. Enhancing AI self-efficacy ensures that transparency functions as a practical resource rather than merely a technical feature.

5.5.Conditional influence on employee innovation

Moderated mediation analysis revealed that the indirect effect of AI transparency on innovation via psychological safety was contingent on AI self-efficacy. The effect was strongest at high self-efficacy, significant at average levels, and negligible at low levels. This pattern highlights that transparency alone does not automatically produce favorable outcomes. Its benefits emerge when employees possess sufficient capability to interpret and act upon AI-related information.

This finding refines the common assumption that increased transparency is universally beneficial. Instead, its effect is context-dependent, emerging when transparency interacts with employee capability beliefs. Moreover, the SME context in China, characterized by hierarchical awareness and interpersonal caution, may heighten employees' sensitivity to whether AI processes are interpretable and discussable, further amplifying the role of transparency in fostering psychological safety and innovation.

5.6.Theoretical contributions

This research advances theoretical understanding in three main areas:

1.Broadening the Scope of AI Transparency Studies: Unlike prior work that primarily examined trust and acceptance, this study connects transparency to concrete employee behaviors, demonstrating its relevance for organizational outcomes.

2.Highlighting Psychological Safety as a Conduit: By showing that psychological safety mediates the relationship between transparency and innovation, the study illustrates how technology features influence employee behavior indirectly through perceptions of safety, bridging human–AI interaction and organizational behavior theories.

3.Elucidating the Role of AI Self-Efficacy: Introducing self-efficacy as a moderating factor clarifies under what conditions and for which employees transparency is most impactful, thereby extending social cognitive theory to AI-driven workplace settings.

Overall, these contributions bridge AI explainability research and workplace behavior studies, emphasizing that technology design and human factors jointly influence organizational outcomes.

5.7.Practical implications

The findings have several practical implications for AI implementation:

1.Prioritize Usable Transparency: Organizations should ensure that explanations and outputs are interpretable and actionable, not just technically available. Interfaces should enable employees to trace decision logic without advanced expertise.

2.Develop AI Self-Efficacy: Training, scenario-based exercises, and guided practice can enhance employees' confidence in using AI, ensuring that transparency cues translate into improved safety and innovation.

3.Foster Psychological Safety: Transparent systems alone cannot generate safe work environments. Managers should legitimize questioning, experimentation, and constructive challenge to AI outputs, integrating system design, capability development, and supportive culture.

5.8.Limitations and future research

This study has several limitations:

1.Cross-Sectional Design: Causal inferences are limited. Longitudinal or experimental studies could confirm temporal ordering among AI transparency, psychological safety, and innovation.

2.Self-Report Measures: Single-source data may inflate relationships. Future research could include supervisor ratings or objective innovation indicators.

3.Context Specificity: Knowledge workers in Chinese SMEs; results may differ in large firms, other industries, or cross-cultural contexts.

4.Role Characteristics: Findings pertain to knowledge-intensive work; routine or standardized roles may exhibit different relationships.

5.Cultural Sensitivity: Hierarchical and interpersonal norms in China may influence perceptions of transparency and safety, necessitating cross-national validation.

Future studies should adopt multi-level, cross-cultural, and multi-industry designs to assess boundary conditions under which AI transparency effectively supports innovation.

6.Conclusion

The present study explored how employees' perceptions of AI transparency relate to their innovative behavior, emphasizing the mediating role of psychological safety and the moderating influence of AI self-efficacy. Results show that transparency supports innovation both directly and indirectly, with the indirect effect via psychological safety being stronger among employees who have greater confidence in using AI.

The results underscore that the organizational value of transparency is not determined solely by technical availability but by employees' ability to interpret, apply, and engage with AI in psychologically safe environments. Transparency is most meaningful when it is understandable, usable, and embedded within supportive work contexts that encourage experimentation and questioning.

By linking AI transparency to employee innovation, the study extends prior research beyond trust and acceptance to behaviorally relevant outcomes, highlighting the interplay between technological attributes and human factors. Practically, organizations should combine transparent AI system design with interventions to enhance employee capability and cultivate psychologically safe environments, thereby fostering sustainable innovation in AI-enabled workplaces.

Author contributions Qian Li: Conceptualization, Methodology, Investigation, Data curation, Formal analysis, Writing – original draft. Peilin Li: Methodology, Formal analysis, Visualization, Writing – review & editing. Chan Sai Keong: Supervision, Writing – review & editing. All authors have read and approved the final version of the manuscript.

Funding No funding was received for this work.

Data availability The data that support the findings of this study are available from the corresponding author upon reasonable request.

Declarations

Conflict of interest The authors declare no conflict of interest.

References

[1] KELLOGG K C, VALENTINE M A, CHRISTIN A. Algorithms at work: The new contested terrain of control. Academy of Management Annals, 2020, 14(1): 366-410.
[2] SOULAMI H, EL BAROUDI S, VAN DEN HEUVEL M. AI at work and employee wellbeing: A systematic review and future research agenda. Frontiers in Artificial Intelligence, 2024, 7: 1473872.
[3] MCKINSEY & COMPANY. The state of AI in 2023: Generative AI's breakout year. 2023.
[4] YU L, LI Y. Artificial intelligence decision-making transparency and employees' trust: The parallel multiple mediating effect of effectiveness and discomfort. Behavioral Sciences, 2022, 12(5): 127.
[5] ZERILLI J, KNOTT A, MACLAURIN J, et al. Transparency in algorithmic and human decision-making: Is there a double standard? Philosophy & Technology, 2019, 32(4): 661-683.
[6] SHIN D. The effects of explainability and causability on perception, trust, and acceptance: Implications for explainable AI. International Journal of Human-Computer Studies, 2021, 146: 102551.
[7] LIU Z. Explainable machine learning for telecom customer churn prediction and actionable retention strategies. Applied Artificial Intelligence Research, 2026, 2(1).
[8] YUAN F. Innovative application of artificial intelligence algorithms in credit scoring: Taking specific products in the field of consumer finance as the research object. Applied Artificial Intelligence Research, 2025, 1(1).
[9] ZHOU J. Personalized recommendation algorithms in digital media: A review of engagement effects, information cocoons, and marketing implications. Applied Artificial Intelligence Research, 2026, 2(1).
[10] ARRIETA A B, DÍAZ-RODRÍGUEZ N, DEL SER J, et al. Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Information Fusion, 2020, 58: 82-115.
[11] YANG M M, LU Y, COOKE F L. Demystifying AI for the workforce: The role of explainable AI in worker acceptance and management relations. Journal of Management Studies, 2026, 63(2): 438-472.
[12] DAVIS F D. Perceived usefulness, perceived ease of use, and user acceptance of information technology. MIS Quarterly, 1989, 13(3): 319-340.
[13] GUNNING D, AHA D W. DARPA's explainable artificial intelligence (XAI) program. AI Magazine, 2019, 40(2): 44-58.
[14] MILLER T. Explanation in artificial intelligence: Insights from the social sciences. Artificial Intelligence, 2019, 267: 1-38.
[15] JANSSEN O. Job demands, perceptions of effort-reward fairness, and innovative work behaviour. Journal of Occupational and Organizational Psychology, 2000, 73(3): 287-302.
[16] CARMELI A, REITER-PALMON R, ZIV E. Inclusive leadership and employee involvement in creative tasks: The mediating role of psychological safety. Creativity Research Journal, 2010, 22(3): 250-260.
[17] BAER M, FRESE M. Innovation is not enough: Climates for initiative and psychological safety, process innovations, and firm performance. Journal of Organizational Behavior, 2003, 24(1): 45-68.
[18] DOSHI-VELEZ F, KIM B. Towards a rigorous science of interpretable machine learning. arXiv, 2017.
[19] SAMEK W, MONTAVON G, VEDALDI A, et al. (Eds.). Explainable AI: Interpreting, explaining and visualizing deep learning. Berlin: Springer, 2019.
[20] RIBEIRO M T, SINGH S, GUESTRIN C. "Why should I trust you?": Explaining the predictions of any classifier//Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM, 2016: 1135-1144.
[21] KELLOGG K C, VALENTINE M A, CHRISTIN A. Algorithms at work: The new contested terrain of control. Academy of Management Annals, 2020, 14(1): 366-410.
[22] MÖHLMANN M, ZALMANSON L, HENFRIDSSON O, et al. Algorithmic management of work on online labor platforms: When matching meets control. MIS Quarterly, 2021, 45(4): 1999-2022.
[23] NEWMAN A, DONOHUE R, EVA N. Psychological safety: A systematic review of the literature. Human Resource Management Review, 2018, 28(3): 521-535.
[24] LIANG B, WANG Y, HUO W, et al. Algorithmic control as a double-edged sword: Its relationship with service performance and work well-being. Journal of Business Research, 2025, 189: 115199.
[25] BANDURA A. Self-efficacy: The exercise of control. New York: W. H. Freeman, 1997.
[26] COMPEAU D R, HIGGINS C A. Computer self-efficacy: Development of a measure and initial test. MIS Quarterly, 1995, 19(2): 189-211.
[27] JIMMIESON N L, TERRY D J, CALLAN V J. A longitudinal study of employee adaptation to organizational change: The role of change-related information and change-related self-efficacy. Journal of Occupational Health Psychology, 2004, 9(1): 11-27.
[28] TIERNEY P, FARMER S M. Creative self-efficacy: Its potential antecedents and relationship to creative performance. Academy of Management Journal, 2002, 45(6): 1137-1148.
[29] EDMONDSON A. Psychological safety and learning behavior in work teams. Administrative Science Quarterly, 1999, 44(2): 350-383.
[30] SCOTT S G, BRUCE R A. Determinants of innovative behavior: A path model of individual innovation in the workplace. Academy of Management Journal, 1994, 37(3): 580-607.
[31] PODSAKOFF P M, MACKENZIE S B, LEE J Y, et al. Common method biases in behavioral research: A critical review of the literature and recommended remedies. Journal of Applied Psychology, 2003, 88(5): 879-903.
[32] KOCK N. Common method bias in PLS-SEM: A full collinearity assessment approach. International Journal of e-Collaboration, 2015, 11(4): 1-10.
[33] SHMUELI G, SARSTEDT M, HAIR J F, et al. Predictive model assessment in PLS-SEM: Guidelines for using PLSpredict. European Journal of Marketing, 2019, 53(11): 2322-2347.
[34] HAIR J F, HULT G T M, RINGLE C M, et al. A primer on partial least squares structural equation modeling (PLS-SEM) (2nd ed.). Thousand Oaks: SAGE, 2019.
[35] HENSELER J, RINGLE C M, SARSTEDT M. A new criterion for assessing discriminant validity in variance-based structural equation modeling. Journal of the Academy of Marketing Science, 2015, 43(1): 115-135.
[36] PREACHER K J, RUCKER D D, HAYES A F. Addressing moderated mediation hypotheses: Theory, methods, and prescriptions. Multivariate Behavioral Research, 2007, 42(1): 185-227.
[37] HAYES A F. Introduction to mediation, moderation, and conditional process analysis: A regression-based approach (2nd ed.). New York: Guilford Press, 2018.