Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Research on Algorithm Aversion in the Field of Academic Assessment —An Explanation Based on Attribution Theory

Ru Liu

Jining Polytechnic

Changlong Liu

Changlong.Liu@hotmail.com

Qufu Normal University

Article

Keywords: academic assessment, algorithm aversion, attribution theory, algorithm failure

Posted Date: November 26th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11474203/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Abstract

As artificial intelligence is applied ever more deeply in educational settings, algorithm aversion has become an important factor affecting the application of automated algorithmic assessment. Through a quasi-experimental design embedded in a real academic assessment setting, this study examined the phenomenon and the psychological mechanisms of algorithm aversion in the academic domain. The results indicate: (1) students show a significant algorithm aversion with respect to automated academic assessment; (2) perceived stability mediates the relationship between the evaluative agent and algorithm aversion; (3) perceived controllability significantly moderates the mediating path through which the evaluative agent influences algorithm aversion via perceived stability, and the moderated mediation effect is established. Specifically, when participants perceive that the algorithm has low controllability over assessment outcomes, the gap between the algorithm and the human teacher in stability ratings is largest, and the algorithm aversion triggered through the stability path is strongest; when participants perceive controllability to be high, the gap narrows and the indirect effect weakens until it disappears. These findings support self-serving bias theory, revealing that individuals tend to attribute algorithmic errors to the algorithm's inherent stable flaws while being more tolerant of human errors. The findings provide an attribution-based explanation of the boundary conditions and psychological mechanisms of algorithm aversion in academic assessment, and offer practical implications for reducing algorithm aversion in the process of automated academic assessment.

1. Introduction

As generative artificial intelligence is applied ever more deeply in educational settings, algorithms are increasingly taking part in people's decision-making processes. On the one hand, algorithms can break through the limitations of human decision making, leading people to place high hopes on improving decision quality through algorithmic decision making (Burton et al., 2020); on the other hand, people also harbor concerns about algorithmic decision making, believing that letting machines make decisions in place of human beings may lead to catastrophic consequences (Bostrom, 2014). Normally, human beings as decision makers are easily disturbed by factors such as prior experience (Kahneman & Tversky, 1972) and emotion (Lerner et al., 2015), whereas algorithmic decision making processes information through the consistent application of preset rules, is not disturbed by factors such as fatigue and emotion (Barnes et al., 2015), and is fast in decision making (Bonnefon et al., 2016), strong in its capacity to encompass information (Blair & Saffidine, 2019), and produces outcomes that are more objective and fair (Andrews et al., 2006). However, psychologists such as Dietvorst (Dietvorst et al., 2014) found in a series of experiments that even when the predictive performance of an algorithm is superior to that of human beings, people still prefer human judgment to algorithmic judgment; this phenomenon is called algorithm aversion. Interestingly, in subsequent research on algorithmic decision making, researchers found that people show two diametrically opposite attitudes toward algorithms in different situations, namely algorithm aversion and algorithm appreciation (Algorithm Appreciation; Logg et al., 2019). In contrast to algorithm aversion, algorithm appreciation refers to the fact that people are sometimes more inclined to

trust and accept the decisions or recommendations of algorithms, believing that algorithms are more reliable and fairer than human beings.

At present, automated academic assessment has become a key application scenario of artificial intelligence empowering education (González-Calatayud, V., 2021), injecting new momentum into the reform of learning assessment (Lu Xiu, 2025). Automated academic assessment is not merely an issue of technological application; it is also a complex cognitive process involving students' self-concept and psychological attribution. Students' attitude of acceptance toward algorithms is an important prerequisite for whether automated assessment can be effectively implemented (Lyons, J.B., 2019). Therefore, conducting in-depth research on attitudes toward algorithms in the field of academic assessment and systematically revealing the formation mechanisms of people's attitudes toward algorithms will, to a certain extent, affect the effectiveness of artificial intelligence's empowerment of education in the future.

2. Literature Review and Research Hypotheses

2.1 Literature Review

Early research on algorithm aversion mostly focused on the characteristics of the algorithmic system itself, such as the "black box" attribute of the decision-making process (Zhao Yijun et al., 2024), lack of transparency (Ananny & Crawford, 2018), and poor explainability (Bonezzi et al., 2022). This line of research, which starts from the perspective of the algorithm, assumes that aversion can be alleviated as long as the predictive accuracy and transparency of the algorithm are improved. However, a growing body of empirical evidence shows that even when algorithmic performance is clearly superior to that of human beings, people still tend to reject algorithmic decisions (Dietvorst et al., 2014, 2020). This has prompted scholars to turn their attention to the cognitive and motivational perspectives of the human subject and to examine the role of attribution errors, self-efficacy, and defensive psychology in the formation of algorithm aversion (Luo Yingyu, 2023).

From the perspective of attribution theory, people show an asymmetry in attribution toward algorithmic errors and human errors. Dietvorst et al. (2014) pointed out that people tend to attribute algorithmic errors to their stable, essential flaws, such as the algorithm's lack of common sense and contextual understanding, while attributing similar human errors to accidental, temporary factors, such as a momentary lapse of attention. According to self-serving bias theory, when individuals face assessment outcomes that carry self-threatening properties, they have a defensive motivation to attribute success internally (to their own ability or effort) and to attribute undesirable outcomes externally (to the external environment). When the evaluative agent is an algorithm, students tend to judge its errors as long-term and unchangeable intrinsic limitations; this attribution pattern is highly likely to trigger algorithm aversion, thereby manifesting as support for the human decision maker.

Weiner's attributional process theory (Weiner B, 1985) holds that attributional processing is not completed in one step; rather, it is a progressive process comprising primary appraisal and secondary appraisal. The dual-process theory of the social cognition school (Evans, 2008; Kahneman, 2011), by contrast, holds that

the attribution process is the result of a dynamic interaction between fast, automatic intuitive processing and slow, controlled analytical reasoning. Among these, heuristics are a major form of intuitive processing. In human-machine interaction scenarios, people hold very high stability expectations of algorithmic systems and generally maintain a cognitive schema based on the "machine heuristic" (Sundar, 2020): because algorithms are free from emotional interference, fatigue constraints, and subjective bias, their decisions should be highly consistent, precise, and procedurally rule-bound. Once this perfection schema encounters an expectation violation, it easily gives rise to algorithm aversion. Therefore, perceived stability is the key mediating variable that carries the chain from "collapse of the cognitive schema → aversive attitude."

Through two experiments, Jones-Jang et al. (Jones-Jang et al., 2023) revealed two opposing mechanisms through which perceived controllability acts on algorithm aversion: when users perceive the algorithm as having high controllability, the frustration brought by algorithmic failure intensifies aversion; when users perceive the algorithm as having low controllability over adverse consequences (i.e., perceived controllability is low), people instead show greater tolerance toward the algorithm and give more positive evaluations of its fairness.

In the three-dimensional structure of attribution, locus of causality, stability, and controllability are three mutually independent cognitive dimensions. Among them, locus of causality answers "does the cause reside in oneself or in the outside," stability answers "is the performance consistently like this," and controllability answers "is the outcome controlled by that agent." Because research on algorithm aversion needs to create a situation of algorithmic decision making and to let participants believe that an algorithm or a human being made the decision, research on algorithm aversion generally does not consider the locus of causality dimension. Existing research has separately examined the roles of stability and controllability in the formation mechanism of algorithm aversion, but there is still a lack of systematic integrative research on how the two work together in the formation of aversive attitudes. Therefore, this study focuses on the process mechanisms through which human perceptions of the stability and controllability of algorithms (perceived stability and perceived controllability) operate in the formation of algorithm aversion. Specifically, this study takes perceived stability as the psychological mediating path through which the evaluative agent influences algorithm aversion, and further constructs a moderated mediation explanatory model in which perceived controllability moderates path a.

2.2 Construction of the First-Stage Moderated Mediation Model and Research Hypotheses

H1: The phenomenon of algorithm aversion in the field of automated academic assessment is significant.

H2: Perceived stability (M) mediates the relationship between the evaluative agent (X) and algorithm aversion (Y).

H3: Perceived controllability (W) significantly moderates the effect of the evaluative agent (X) on perceived stability (M) (i.e., it moderates the first stage of the mediating process: path a).

H4: Perceived controllability moderates the mediating effect of "evaluative agent → perceived stability → algorithm aversion"; that is, a first-stage moderated mediation effect exists.

It is worth noting that most existing studies rely on hypothetical scenario-based text tests (e.g., Du Xiufang et al., 2023; Jones-Jang et al., 2023). Because such designs lack real stakes and decision-making pressure, they are prone to hypothetical bias, which limits the ecological validity of their conclusions (Bombín-González et al., 2014). To this end, this study adopts a quasi-experimental design and embeds algorithmic assessment into college students' real oral English academic assessment, in order to reveal the deep attributional psychological mechanism of algorithm aversion in highly self-relevant scenarios.

3. Experimental Procedure

3.1 Method

3.1.1 Participants

Before the study was conducted, GPower 3.1.9.7 was used to calculate that, for an independent-samples t-test with a medium effect size (d = 0.5; Cohen, 1992) and a significance level of α = 0.05, the minimum sample size based on a statistical power of 0.9 was 172. This study adopted a real-scenario quasi-experimental design and selected a total of 376 students from 10 parallel classes at a certain university; all participants took part in the experiment voluntarily and gave informed consent. **All methods were carried out in accordance with relevant guidelines and regulations. All experimental protocols were approved by the Ethics Committee of Jining Polytechnic**. Informed consent was obtained from all participants. Using the average scores of the three most recent English proficiency tests as the basis, ANOVA was used to test differences in English proficiency among the classes; P > 0.05 was found, so it was concluded that there were no significant differences in English proficiency among the classes. After excluding 4 questionnaires that failed the attention check and 17 questionnaires that failed the manipulation check, a total of 355 valid questionnaires were obtained for statistical analysis (valid response rate 94.4%), including 166 males (46.7%) and 189 females (53.2%), with a mean age of M = 21 years, SD = 1.24 years.

3.1.2 Experimental Design and Implementation

We adopted a one-factor, two-level between-subjects experimental design and randomly assigned the participants, by class, to a human group and an algorithm group, with 187 participants in the human group and 168 in the algorithm group. During the experiment, participants were asked to report the permissibility of the assessment result, their degree of liking, and their willingness to use it, so as to compare differences in attitudes toward human assessment and algorithmic assessment and to test whether algorithm aversion exists in the field of academic assessment.

This study used an oral English assignment as the research task, mainly because an oral English assignment perfectly satisfies the three characteristics of subjectivity, black-box nature, and high self-relevance, making it an experimental material that can effectively trigger algorithm aversion. All

participants were asked to give an oral English presentation on the theme "Now Is the Best Time to Start" (time limit: 2 minutes ± 30 seconds). To ensure consistency of experimental conditions, the task instructions were read aloud on site by the same experimenter (female, Putonghua proficiency Level 1-B) using a standardized script, and all participants were required to complete the task independently without the aid of any other tools. Next, participants' responses were audio-recorded, and two experimenters with similar educational backgrounds scored them independently, with the scores averaged. The results showed that the difference in oral performance scores among classes did not reach statistical significance ($H(9) = 12.15$, $p = 0.206$), indicating that the differences in oral performance among students in different classes were not significant, and that the ratings of the two experimenters had good consistency (ICC = 0.79).

According to the view of Glikson E. and Woolley A. W. (Glikson E, Woolley A W., 2020), algorithms can be divided into embedded algorithms, algorithm-empowered virtual robots, and algorithm-empowered physical robots. The algorithm used in this study refers to an algorithm-empowered virtual robot. The task prompt given to the teacher group was: The score given by the teacher is ; please complete the questionnaire according to your actual situation. The task prompt given to the algorithm group was: The assessment score obtained by using Automated Oral English Scoring 2.0 developed by Learning Agency Lab of a certain country is ; please complete the questionnaire according to your actual situation. Subsequently, participants were asked to respond on the questionnaire. The questionnaire consisted of 7 items in total, including one attention-check item and one manipulation-check item.

Drawing on existing measurement instruments (Campbell, 1999; Jones-Jang et al., 2023; Zhao Yijun, 2024), we asked participants about their attitudes toward the human teacher and the algorithm through five questions; all items were rated on a 7-point scale:

Permissibility: (Q1) Do you think the score given by the teacher (the algorithm AoES 2.0) is appropriate?

Liking: (Q2) To what extent do you agree with the evaluation made by the teacher (the algorithm AoES 2.0)?

Willingness to Use: (Q3) To what extent do you support the teacher (the algorithm "AoES 2.0") in carrying out subsequent assessment work?

Perceived Stability: (Q4) Do you think the evaluation criteria of the teacher (the algorithm AoES 2.0) are consistent throughout?

Perceived Controllability: (Q5) How much control do you think the teacher (the algorithm AoES 2.0) has in determining the score?

3.1.3 Attention Check and Manipulation Check

An attention-check item was inserted among the items: For this item, please select 5. Participants who did not select 5 were regarded as having failed the attention check. A manipulation-check item was inserted among the items: Was this assessment result given by a teacher (an algorithm)? Participants

whose answer was inconsistent with the instructions were regarded as having failed the manipulation check.

3.2 Results

3.2.1 Test of the Main Effect

An independent-samples t-test was conducted with the evaluative agent (human teacher group = 1; algorithm group = 2) as the independent variable and the participants' responses to the three items (Q1-Q3) as the dependent variables; the results are shown in Table 1. (Permissibility, liking, and willingness to use)

Table 1
Results of the Independent-Samples t-Test

Items	Homogeneity of Variance	Levene's Test		t-Test			
		F	Significance	t	Mean Difference	Sig. (Two-Tailed)	95% Confidence Interval of the Difference
Q1: Permissibility	Equal variances assumed	.971	.325	3.595	.488	.000	.221 ~ 0.755
Q2: Liking	Equal variances assumed	.005	.945	2.883	.451	.004	.143 ~ .758
Q3: Willingness to Use	Equal variances assumed	.062	.804	8.874	1.039	.000	.809 ~ 1.269

The t-test results show that the phenomenon of algorithm aversion in the field of academic assessment is significant, consistent with the conclusions of research in other domains (Zhao Yijun, 2024; Chun Jinseok S, 2024): participants agreed more with the assessment results produced by human teachers and considered the scores given by human teachers to be fairer; with respect to assessment consistency, they regarded the assessments of human teachers as more stable; and they supported human teachers in carrying out subsequent assessment work. H1 was verified.

3.2.2 Test of the Mediating Effect of Perceived Stability

3.2.2.1 Multiple Regression Analysis and Collinearity Test

An internal consistency test was conducted on the responses to the items on the three dimensions of algorithm aversion (Cronbach's α = 0.794). The results show that the internal consistency of the three items is good, so the average value can be used as the overall index of algorithm aversion.

After centering, a linear multiple regression analysis was conducted with the evaluative agent, perceived stability, and perceived controllability as independent variables and algorithm aversion as the dependent variable. The collinearity test results show that the tolerance of the three independent variables was all > 0.6 (range 0.672–0.909), and the variance inflation factors (VIF) were all < 1.5 (range 1.100-1.488), far below the critical value of 5, indicating that there is no multicollinearity problem among the variables and that the regression results are reliable.

The results show that the overall model is significant ($F(3,351) = 141.36$, $p < .001$), with an adjusted $R^2 = .543$; the three variables jointly explain 54.3% of the variance in algorithm aversion. The variable of evaluative agent ($\beta = -.078$, $p = .038$) has a weak negative effect, consistent with the previous t-test results; perceived stability ($\beta = .588$, $p < .001$) has the strongest positive predictive effect on algorithm aversion, indicating that the more individuals believe the stability of algorithmic decision making to be high, the more pronounced their aversive feeling becomes; perceived controllability ($\beta = .196$, $p < .001$) likewise positively predicts aversion, that is, the greater the controllability over the outcome, the stronger the algorithm aversion.

Table 2
Results of the Regression Analysis with Algorithm Aversion as the Dependent Variable

	Unstandardized Coefficients		Standardized Coefficients	t	Significance	Collinearity Statistics
	B	Standard Error	β			Tolerance
(Constant)	2.081	.272		7.640	.000	
Evaluative Agent	− .176	.085	− .078	-2.078	.038	.909
Perceived Stability	.522	.039	.588	13.412	.000	.672
Perceived Controllability	.172	.037	.196	4.592	.000	.709

It should be noted that Table 2 presents the results of entering the three predictor variables into the regression equation simultaneously, with the aim of testing the independent predictive power of each variable and the collinearity situation; the subsequent PROCESS analysis, by contrast, is based on the moderated mediation model (Model 7) for path coefficient estimation. The two differ in statistical principle, and therefore there are slight differences in the coefficient estimates β.

3.2.2.2 Test of the Mediating Path and Model Estimation

According to the literature analysis and the regression analysis results, perceived stability and perceived controllability may serve as mediating variables explaining the mechanism through which the independent variable, the evaluative agent, affects algorithm aversion (the dependent variable). Therefore, the PROCESS procedure (Model 4) developed by Hayes (2022) was used to test the mediating effect and

verify this hypothesis. Based on 5000 Bootstrap samples, the results with 95% confidence intervals are as follows:

The model fit indices with algorithm aversion as the outcome variable are $R^2 = 0.52$, $F(2,352) = 190.62$, $p < 0.001$; the direct effect of the evaluative agent on algorithm aversion is -0.19 (SE = 0.09, p = 0.026, 95% CI [-0.37, -0.02]), indicating that after controlling for the mediating variable, the evaluative agent still has a significant negative effect on algorithm aversion. The mediating effect through perceived stability is -0.46 (BootSE = 0.08, 95% CI [-0.63, -0.31]); the confidence interval does not contain 0, indicating that the mediating effect is significant. Specifically, compared with the human teacher, when the evaluative agent is an algorithm, students' perceived stability is significantly reduced (path a is negative), while perceived stability significantly and positively predicts algorithm aversion (path b is positive), thereby producing a significant negative indirect effect (indirect effect = -0.46, BootSE = 0.08, 95% CI [-0.63, -0.31]). The results indicate that perceived stability plays a partial mediating role between the evaluative agent and algorithm aversion; that is, the algorithmic evaluative agent indirectly influences the level of algorithm aversion by lowering individuals' perception of the stability of the assessment. H2 was verified.

3.2.3 Test of the Moderated Mediation Effect

To test whether perceived controllability moderates the first stage of the mediating path "evaluative agent → perceived stability → algorithm aversion" (H3 and H4), the PROCESS macro (Model 7) developed by Hayes (2022) was used for path analysis. The number of Bootstrap samples was set to 5000 and the confidence interval to 95%. The evaluative agent (X; human group = 1, algorithm group = 2) was the independent variable, perceived stability (M) the mediating variable, perceived controllability (W) the moderating variable, and algorithm aversion (Y) the dependent variable. All continuous variables were mean-centered before analysis.

3.2.3.1 Test of the Moderating Effect

A regression analysis was conducted with perceived stability as the dependent variable and the evaluative agent, perceived controllability, and their interaction term as independent variables; the results are shown in Table 3. The overall model is significant ($F(3,351) = 60.87$, $p < 0.001$), with an explanatory power of $R^2 = 0.3422$. The evaluative agent has a significant negative predictive effect on perceived stability ($\beta = -0.2028$, B = -0.4900, SE = 0.1122, $p < 0.001$), indicating that, compared with the human teacher, students rate the stability of the algorithm lower. The main effect of perceived controllability is not significant ($\beta = 0.0638$, B = 0.1467, SE = 0.1319, p = 0.2667). More importantly, the interaction term between the evaluative agent and perceived controllability significantly predicts perceived stability ($\beta = 0.1645$, B = 0.2473, SE = 0.0892, p = 0.0058), indicating that perceived controllability significantly moderates the effect of the evaluative agent on perceived stability. H3 was verified.

Table 3
Results of the Regression Analysis with Perceived Stability as the Dependent Variable

Predictor	B	SE	β	t	p	95% CI
Constant	0.7570	0.1738		4.3547	< 0.001	[0.4151, 1.0988]
Evaluative Agent (X)	-0.4900	0.1122	-0.2028	-4.3687	< 0.001	[-0.7106, -0.2694]
Perceived Controllability (W)	0.1467	0.1319	0.0638	1.1126	0.2667	[-0.1126, 0.4061]
X × W	0.2473	0.0892	0.1645	2.7732	0.0058	[0.0719, 0.4226]

Note: B is the unstandardized coefficient and β is the standardized coefficient; CI is the bias-corrected Bootstrap confidence interval.

To reveal the specific pattern of the interaction term, we further analyzed the conditional effect of the evaluative agent on perceived stability at different levels of perceived controllability (16th, 50th, and 84th percentiles). The results are shown in Table 4:

Table 4
Conditional Effect of the Evaluative Agent on Perceived Stability

Level of Perceived Controllability	Effect	SE	t	p	95% CI
Low Controllability (16th Percentile)	-0.8016	0.1543	-5.1964	< 0.001	[-1.1049, -0.4982]
Medium Controllability (50th Percentile)	-0.5543	0.1133	-4.8941	< 0.001	[-0.7770, -0.3315]
High Controllability (84th Percentile)	− .0597	0.1965	-0.3041	0.7612	[-0.4461, 0.3266]

The conditional effect analysis indicates that when students perceive the algorithm as having low controllability over assessment outcomes, the negative effect of the evaluative agent on perceived stability is largest (Effect = -0.8016, p < 0.001); as perceived controllability rises, this negative effect gradually weakens; when perceived controllability is at a relatively high level, the effect of the evaluative agent on perceived stability is no longer significant (Effect = -0.0597, p = 0.7612). The results indicate that perceived controllability narrows the gap between the algorithm and the human being in perceived stability: the more students believe that the algorithm can control the assessment outcome, the smaller the difference in stability ratings between the algorithm and the human teacher becomes.

3.2.3.2 Test of the Moderated Mediation Effect

A regression analysis was conducted with algorithm aversion as the dependent variable and the evaluative agent and perceived stability as independent variables; the results show (see Table 5) that the overall model is significant ($F(2, 352) = 190.62$, $p < 0.001$), with an explanatory power of $R^2 = 0.5199$. The evaluative agent has a significant negative direct effect on algorithm aversion ($\beta = -0.0805$, $B = -0.1945$, $SE = 0.0871$, $p = 0.0262$), and perceived stability has an extremely significant positive predictive effect on algorithm aversion ($\beta = 0.6132$, $SE = 0.0344$, $p < 0.001$), indicating that perceived stability is a strong predictor of algorithm aversion.

Table 5
Results of the Regression Analysis with Algorithm Aversion as the Dependent Variable

Predictor	**B**	**SE**	**β**	**t**	**p**	**95% CI**
Constant	5.9164	0.1349		43.8458	< 0.001	[5.6510, .1818]
Evaluative Agent (X)	-0.1945	0.0871	-0.0805	-2.2322	0.0262	[-0.3659, – .0231]
Perceived Stability (M)	0.6132	0.0344	0.6164	17.8432	< 0.001	[0.5456, 0.6808]

We further tested the conditional indirect effects at different levels of perceived controllability and the index of moderated mediation; the results are shown in Table 6.

Table 6
Results of the Moderated Mediation Test (PROCESS Model 7)

Level of Perceived Controllability	**Conditional Indirect Effect**	**Boot SE**	**95% Bootstrap CI**
Low Controllability (16th Percentile)	-0.4915	0.1228	[-0.7258, -0.2508]
Medium Controllability (50th Percentile)	-0.3399	0.0771	[-0.4912, -0.1851]
High Controllability (84th Percentile)	-0.0366	0.1245	[-0.2839, 0.1981]
Index of Moderated Mediation (Index)	0.1516	0.0681	[0.0145, 0.2819]
Note: Number of Bootstrap samples = 5000; CI is the bias-corrected confidence interval.			

The conditional indirect effect analysis indicates that when perceived controllability is low (16th percentile), the indirect effect of the evaluative agent on algorithm aversion through perceived stability is significant (effect size = -0.4915, BootSE = 0.1228, 95% CI [-0.7258, -0.2508]), and the confidence interval does not contain 0; when perceived controllability is at a medium level (50th percentile), the indirect effect is significant (effect size = -0.3399, BootSE = 0.0771, 95% CI [-0.4912, -0.1851]), and the confidence interval does not contain 0; when perceived controllability is high (84th percentile), the indirect effect weakens and is no longer significant (effect size = -0.0366, BootSE = 0.1245, 95% CI [-0.2839, 0.1981]), and the confidence interval contains 0.

The index of moderated mediation is 0.1516 (BootSE = 0.0681), with a 95% Bootstrap confidence interval of [0.0145, 0.2819], which does not contain 0, indicating that the first-stage moderated mediation effect is established, as shown specifically in Fig. 1. H4 was verified.

3.2.3.3 Summary

Taken together, the above results show that: perceived stability plays a partial mediating role between the evaluative agent and algorithm aversion (H2 supported); perceived controllability significantly moderates the effect of the evaluative agent on perceived stability (H3 supported), and this moderating effect is further transmitted to the mediating path, forming a first-stage moderated mediation effect (H4 supported). Specifically, when students perceive the algorithm as having low controllability over assessment outcomes, the gap between the algorithm and the human teacher in stability ratings is largest, and the algorithm aversion generated through the stability path is also strongest; when perceived controllability is high, this gap narrows, and the indirect effect of algorithm aversion weakens accordingly until it disappears.

3.2.4 Supplementary Analysis: Test of the Mediating Effect of Perceived Controllability

Similarly, the PROCESS procedure (Model 4) developed by Hayes (2022) was used, with perceived control as the mediating variable, to conduct a mediating effect analysis of the relationship between the evaluative agent and algorithm aversion. Based on 5000 Bootstrap samples, the results with 95% confidence intervals are as follows:

The explanatory power of the model with algorithm aversion as the outcome variable is $R^2 = 0.32$, $F(2,352) = 80.96$, $p < 0.001$. The path from the evaluative agent to perceived control shows that the evaluative agent has a significant negative predictive effect on perceived control ($\beta = -0.51$, SE = 0.13, $p = 0.0002$, 95% CI [-0.77, -0.25]). The path from perceived control to algorithm aversion shows that perceived control has a significant positive predictive effect on algorithm aversion ($\beta = 0.43$, SE = 0.04, $p < 0.001$, 95% CI [0.35, 0.51]).

The direct effect of the evaluative agent on algorithm aversion is -0.44 (SE = 0.10, $p < 0.001$, 95% CI [-0.64, -0.24]), indicating that after controlling for the mediating variable, the evaluative agent still maintains a significant negative effect on algorithm aversion; the mediating effect through perceived control is -0.22 (SE = 0.06, 95% CI [-0.35, -0.10]), and its confidence interval does not contain 0, indicating that the mediating effect is significant. Specifically, the mediating path of lowering perceived control and thereby reducing algorithm aversion is established.

The results show that evaluation has a significant negative predictive effect on perceived control ($\beta = -0.51$, SE = 0.13, $p = 0.0002$), and that perceived control has a significant positive predictive effect on algorithm aversion ($\beta = 0.43$, SE = 0.04, $p < 0.001$). The indirect effect is -0.22 (BootSE = 0.06, 95% CI [-0.35, -0.10]); the confidence interval does not contain 0, indicating that the mediating effect of perceived control is significant.

This result is consistent with the findings of Jones-Jang et al. (Jones-Jang et al., 2023) and verifies the existence of the "perceived control path" in educational settings. However, because this path is not a core hypothesis of this study, it is reported only as a supplement; in the discussion below, we will conduct a theoretical analysis in light of this result.

4. Discussion

Through a quasi-experiment embedded in a real academic assessment setting, this study examined the phenomenon of algorithm aversion among college students and its psychological mechanisms based on attribution theory. The results not only confirm the phenomenon of algorithm aversion in the field of academic assessment, but also reveal more deeply the dual mechanisms (namely the mediating and moderating roles) of perceived stability and perceived controllability therein. This study finds that individuals' aversion to algorithms is not a simple rejection, but a complex cognitive and motivational process driven by attribution. In what follows, we combine the results of this study with the existing literature to discuss in an integrated manner the theoretical contributions, practical implications, and future directions of the research findings.

4.1 Implications for Carrying Out Automated Academic Assessment

The results of this study have certain practical significance for the application of artificial intelligence in the field of education.

First, improve the transparency and explainability of algorithms. In the cognitive conceptions of ordinary people, algorithms possess a typical "black box" characteristic, which makes it difficult for them to perceive the transparency and controllability of algorithms (Bonezzi et al., 2022). In the context of academic assessment, the "black box" characteristic of algorithms—that is, the opacity and inexplicability of their decision-making process—constitutes precisely this kind of "heterogeneity" that people can neither adapt to nor enter into dialogue with. This heterogeneity deprives students of a sense of understanding and participation, and becomes the root cause of algorithm aversion. Therefore, when designing academic assessment algorithms, one should not merely output a cold score; rather, one should provide clear assessment rationales, process feedback, and improvement suggestions, so as to enhance students' acceptance of algorithmic assessment.

Second, manage user expectations. Since algorithm aversion originates from the violation of high expectations, developers and users of educational technology should consciously manage students' expectations regarding algorithmic capabilities. Clearly informing students of the capability boundaries of algorithms, the errors that may exist, and their auxiliary positioning helps to establish a more rational and tolerant environment of human-machine collaborative assessment, and to alleviate at the source the strong antipathy caused by the collapse of the perfection schema.

Third, adopt a hybrid assessment model. At the initial stage of introducing automated assessment, a hybrid model of "algorithm-led + human-in-the-loop" can be adopted, which both makes use of the efficiency of algorithms and retains the advantages of human teachers in handling complex, contextual, and creative tasks, while at the same time giving students a channel to appeal and to seek human explanation; this can effectively buffer the negative experience brought about by algorithm aversion.

4.2 Attribution Theory as an Effective Explanatory Framework for the Phenomenon of Algorithm Aversion

The most important theoretical contribution of this study lies in providing, from the perspective of attribution theory, an analytical framework with deep explanatory power for the phenomenon of algorithm aversion. Although previous studies have repeatedly confirmed the existence of algorithm aversion (Dietvorst et al., 2014; Burton et al., 2020), with respect to the core question of "why people's tolerance of algorithmic errors is far lower than their tolerance of human errors," most have remained at the descriptive level (Zhao Yijun et al., 2024), lacking a mechanistic explanation of the underlying attribution process. Through a quasi-experimental design embedded in a real academic assessment setting, this study introduces Weiner's attributional process theory (Weiner B, 1985) and self-serving bias theory (Miller & Ross, 1975) into algorithm aversion research, and reveals three key psychological mechanisms:

First, stability attribution is the core driving force of algorithm aversion. This study finds that both the effect of the evaluative agent on perceived stability ($\beta = -0.49$, $p < 0.001$) and the predictive effect of perceived stability on algorithm aversion ($\beta = 0.61$, $p < 0.001$) are highly significant, and the mediating effect is established. This means that students' aversion to algorithms does not stem from the algorithm itself, but from the algorithm being attributed as an assessment tool that is "inherently unstable." As machine heuristic theory (Sundar, 2020) points out, people hold a cognitive schema of the "perfect performance" of algorithms; once the algorithm's performance fails to meet expectations—even if it is comparable to the level of a human teacher—its errors are attributed to essential flaws ("the algorithm is unstable," "the machine is unreliable") rather than to situational factors ("this time it was just an accident").

Second, self-serving bias activates defensive attribution motivation. In the highly self-relevant scenario of academic assessment, assessment outcomes are directly related to students' academic self-concept (Pekrun, 2006). When the evaluative agent is an algorithm, attributing an unsatisfactory assessment result to the "unreliability" of the algorithm (an external, stable, uncontrollable factor), rather than to one's own insufficient English proficiency (an internal, unstable, controllable factor), constitutes a psychological strategy for protecting self-esteem. This finding is consistent with the conclusions of the meta-analysis of Campbell and Sedikides (1999): the stronger the self-threat, the more inclined individuals are to adopt external attribution in order to maintain a positive self-image.

Third, the attribution process is not a single linear processing, but a dynamic interaction of cognitive and motivational factors. Through the moderated mediation model, this study reveals that perceived controllability, as an independent dimension in the three-dimensional structure of attribution (Weiner B, 1985), not only independently influences algorithm aversion, but also, by moderating the first stage of the

attribution path "evaluative agent → perceived stability," changes the individual's overall attribution pattern. This finding indicates that individuals' attribution toward algorithms is not completed in one step (the static model implied in Jussupow et al., 2020); rather, it goes through a progressive process of primary appraisal (stability judgment) and secondary appraisal (controllability attribution), and the two act jointly on the formation of aversive emotion in a manner of "moderation rather than superposition."

In sum, attribution theory provides a theoretical lens for understanding algorithm aversion that possesses both explanatory power and integrative capacity. It explains not only "why" algorithm aversion occurs, but also reveals "under what conditions it is strengthened or weakened"; this is precisely the theoretical increment that distinguishes this study from existing research on algorithm aversion (Dietvorst et al., 2014, 2020; Jones-Jang et al., 2023).

4.3 Verification and Extension of the Jones-Jang Dual-Path Model

This study not only verifies the mediating path of perceived stability (H2), but also reveals, through the moderated mediation model, the moderating role of perceived controllability in the first stage of the path. This result shows both consistency with and important extension of the dual-path model proposed by Jones-Jang et al. (2023).

Consistent with the findings of Jones-Jang et al., this study confirms the existence of the "stability path": students hold "machine heuristic" expectations of algorithms, and fluctuations in the algorithm's performance are more readily attributed to inherent, stable flaws, thereby triggering stronger algorithm aversion. This study further finds that this attribution process is significantly moderated by perceived controllability, but the direction of the moderation differs from the controllability "absolution effect" of Jones-Jang et al. (Jones-Jang et al., 2023).

Specifically, this study finds that when students perceive the algorithm as having low controllability over assessment outcomes (that is, the algorithm "does not have the final say"), the gap between the algorithm and the human teacher in stability ratings is largest, and the indirect effect generated through the stability path is strongest; whereas when the algorithm is believed to have relatively high controllability, this gap instead narrows and the indirect effect disappears. This pattern suggests that granting the algorithm greater decision-making control may instead change students' "machine heuristic" cognitive schema of the algorithm: when the algorithm is regarded as a "quasi-agent" with substantial decision-making authority, students' harsh judgment of its stability decreases, and the "heterogeneity" of the algorithm as an evaluative agent is partially dissolved. In other words, the stronger the algorithm's "sense of agency," the less it resembles a "cold machine," and the closer students' attribution toward it becomes to their attribution pattern toward human teachers.

This finding provides a new theoretical perspective for understanding the dynamic boundary of human-machine trust: the degree of "anthropomorphism" of an algorithm is not simply the higher the better or the lower the better; rather, it produces complex nonlinear effects across different attribution dimensions.

4.4 Research Limitations and Future Directions

This study still has several limitations, which point the way for future research.

First, this study focuses on the "normal assessment scenario," in which all participants obtained assessment results at a medium level (they neither encountered significantly low ratings nor obtained unusually high ratings). Although this design effectively controlled for the interference of the extremity of the assessment result itself with the attribution process, it also limits the scope of generalization of the research conclusions to the "assessment failure scenario." A highly promising future direction is to manipulate the direction of the assessment result (for example, the algorithm significantly overestimates or significantly underestimates students' true level) and to explore how the attribution mechanism changes under failure scenarios. Based on self-serving bias theory, we speculate that under failure scenarios, students will more strongly activate defensive attribution motivation, externally attributing negative outcomes to the "unfairness" or "error" of the algorithm, so that algorithm aversion is sharply amplified; whereas under success scenarios, the algorithm may be "tolerated" or even appreciated (the algorithm appreciation effect). Verifying this hypothesis would greatly enrich our understanding of the boundary conditions of algorithm aversion.

Second, this study found a counterintuitive moderation pattern: when perceived controllability is high, the gap between the algorithm and the human being in stability ratings instead narrows, and the indirect effect disappears. This finding suggests that the algorithm's "sense of agency" may have changed individuals' attribution schema, but the specific cognitive path of this mechanism is not yet clear. Future research may adopt methods such as the process dissociation procedure or eye-tracking to examine differences in the duration of attributional processing and the allocation of attention under different perceptions of controllability, in order to reveal its micro-level cognitive process.

In addition, this study adopts a cross-sectional design and was unable to capture the dynamic changes in aversion after multiple interactions between students and algorithms. Future research may adopt a longitudinal tracking design to examine whether repeated exposure produces a habituation effect, thereby alleviating the initial algorithm aversion.

Finally, the sample of this study consists entirely of Chinese college students, and the cultural background is single. Under different cultural backgrounds (for example, collectivism vs. individualism), individuals' acceptance of algorithmic authority and their attribution preferences may differ significantly. Cross-cultural comparison will be an important direction for verifying the external validity of the conclusions of this study. In addition, applying attributional retraining to alleviate algorithm aversion and testing its intervention effectiveness is also a frontier topic with both theoretical value and practical significance.

4.5 Conclusion

Grounded in a real educational setting, this study confirms the existence of algorithm aversion in academic assessment and, based on attribution theory, reveals its underlying psychological mechanism: perceived stability is the key mediating variable linking the evaluative agent and algorithm aversion, and

perceived controllability significantly moderates the first stage of this mediating path—when the algorithm is believed to have relatively high controllability over the outcome, the gap between it and the human teacher in stability ratings instead narrows, and the indirect effect weakens until it disappears. This finding indicates that individuals' attitudes toward algorithms are not simple technological acceptance or rejection, but a complex cognitive process driven by attribution and deeply influenced by self-serving bias. This study not only deepens the theoretical understanding of algorithm aversion, but also provides an empirical basis for the responsible design of artificial intelligence assessment systems in the field of education.

Declarations

Funding Information:

2025 Key Project of the Ministry of Education under the National Education Science Planning Program, "Research on the Key Influencing Factors of the Digital Transformation of Vocational Education and Their Mechanisms of Action" (Project No.: DJA250408; Principal Investigator: Lü Jianqiang).

Author Contribution

R.L. and C.L. designed the study, conducted the experiment, analysed the data, and wrote the main manuscript text. All authors reviewed the manuscript.

Data Availability

The datasets used and analysed during the current study are available from the corresponding author on reasonable request.

References

1. Ananny, M. & Crawford, K. Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. *New. Media Soc.* **20**, 973–989 (2018).
2. Andrews, D., Bonta, J. & Wormith, J. S. The recent past and near future of risk and/or need assessment. *Crime. Delinq*. **52**, 7–27 (2006).
3. Barnes, C. M. et al. You wouldn't like me when I'm sleepy: Leaders' sleep, daily abusive supervision, and work unit engagement. *Acad. Manage. J.* **58**, 1419–1437 (2015).
4. Blair, A. & Saffidine, A. AI surpasses humans at six-player poker. *Science* **365**, 864–865 (2019).
5. Bombín-González, I. et al. Ecological validity and multitasking environments in the evaluation of the executive functions. *Rev. Neurol.* **59**, 77–87 (2014).
6. Bonezzi, A., Ostinelli, M. & Melzner, J. The human black-box: The illusion of understanding human better than algorithmic decision making. *J. Exp. Psychol. Gen.* **151**, 2250–2258 (2022).

7. Bonnefon, J. F., Shariff, A. & Rahwan, I. The social dilemma of autonomous vehicles. *Science* **352**, 1573–1576 (2016).
8. Bostrom, N. & Superintelligence *Paths, Dangers, Strategies* (Oxford University Press, 2014).
9. Burton, J. W., Stein, M. K. & Jensen, T. B. A systematic review of algorithm aversion in augmented decision making. *J. Behav. Decis. Mak.* **33**, 220–239 (2020).
10. Calatayud, V. G. & Prendes, P. Roig-Vila, R. Artificial intelligence for student assessment: A systematic review. *Appl. Sci.* **11**, 5467 (2021).
11. Campbell, M. C. Perceptions of price unfairness: Antecedents and consequences. *J. Mark. Res.* **36**, 187–199 (1999).
12. Campbell, W. K. & Sedikides, C. Self-threat magnifies the self-serving bias: A meta-analytic integration. *Rev. Gen. Psychol.* **3**, 23–43 (1999).
13. Chun, J. S. et al. What algorithmic evaluation fails to deliver: respectful treatment and individualized consideration. *Sci. Rep.* **14**, 25996 (2024).
14. Cohen, J. A power primer. *Psychol. Bull.* **112**, 155–159 (1992).
15. Dietvorst, B. J. & Bharti, S. People reject algorithms in uncertain decision domains because they have diminishing sensitivity to forecasting error. *Psychol. Sci.* **31**, 1302–1314 (2020).
16. Dietvorst, B. J., Simmons, J. P. & Massey, C. Algorithm aversion: People erroneously avoid algorithms after seeing them err. *J. Exp. Psychol. Gen.* **144**, 114–126 (2014).
17. Du, X., Wang, J., Li, F. & Wang, Y. Algorithm aversion or algorithm appreciation? The influence of different advice sources on individuals' advice taking. *Chin. J. Clin. Psychol.* **31**, 75–79 (2023). (in Chinese).
18. Evans, J. St. B. T. Dual-processing accounts of reasoning, judgment, and social cognition. *Annu. Rev. Psychol.* **59**, 255–278 (2008).
19. Glikson, E. & Woolley, A. W. Human trust in artificial intelligence: Review of empirical research. *Acad. Manage. Ann.* **14**, 627–660 (2020).
20. Jones-Jang, S. M. & Park, Y. J. How do people react to AI failure? Automation bias, algorithmic aversion, and perceived controllability. *J. Comput. -Mediat Commun.* **28**, zmac023 (2023).
21. Jussupow, E., Benbasat, I. & Heinzl, A. Why are we averse towards algorithms? A comprehensive literature review on algorithm aversion. in Proceedings of the European Conference on Information Systems (2020).
22. Kahneman, D. & Tversky, A. Subjective probability: A judgment of representativeness. *Cogn. Psychol.* **3**, 25–48 (1972).
23. Kahneman, D. & Thinking *Fast and Slow* (Farrar, Straus and Giroux, 2011).
24. Lerner, J. S. et al. Emotion and decision making. *Annu. Rev. Psychol.* **66**, 799–823 (2015).
25. Logg, J. M., Minson, J. A. & Moore, D. A. Algorithm appreciation: People prefer algorithmic to human judgment. *Organ. Behav. Hum. Decis. Process.* **151**, 90–103 (2019).

26. Lu, X. From disembodiment to embodiment: Risks of applying artificial intelligence in learning evaluation and their resolution. *China Educ. Technol.* (6), 102–108 (2025). (in Chinese).
27. Luo, Y. et al. Algorithm aversion in the era of artificial intelligence: A research framework and future prospects. *J. Manag World*. **39**, 205–233 (2023). (in Chinese).
28. Lyons, J. B. et al. in *Trust and human-machine teaming: A qualitative study. in Artificial Intelligence for the Internet of Everything*. 101–116 (eds Lawless, W. F.) (Academic, 2019).
29. Miller, D. T. & Ross, M. Self-serving biases in the attribution of causality: Fact or fiction? Psychol. *Bull* **82**, 213–225 (1975).
30. Pekrun, R. The control-value theory of achievement emotions: Assumptions, corollaries, and implications for educational research and practice. *Educ. Psychol. Rev.* **18**, 315–341 (2006).
31. Sundar, S. S. Rise of machine agency: A framework for studying the psychology of human-AI interaction (HAII). *J. Comput. -Mediat Commun.* **25**, 74–88 (2020).
32. Sundar, S. S. The MIT Press,. The MAIN model: A heuristic approach to understanding technology effects on credibility. in Digital Media, Youth, and Credibility (eds. Metzger, M. J. & Flanagin, A. J.) 73–100 (2008).
33. Weiner, B. An attributional theory of achievement motivation and emotion. *Psychol. Rev.* **92**, 548–573 (1985).
34. Zhao, Y., Xu, L., Yu, F. & Jin, W. Perceived opacity increases algorithm aversion in the workplace. *Acta Psychol. Sin*. **56**, 497–514 (2024). (in Chinese).

Figures

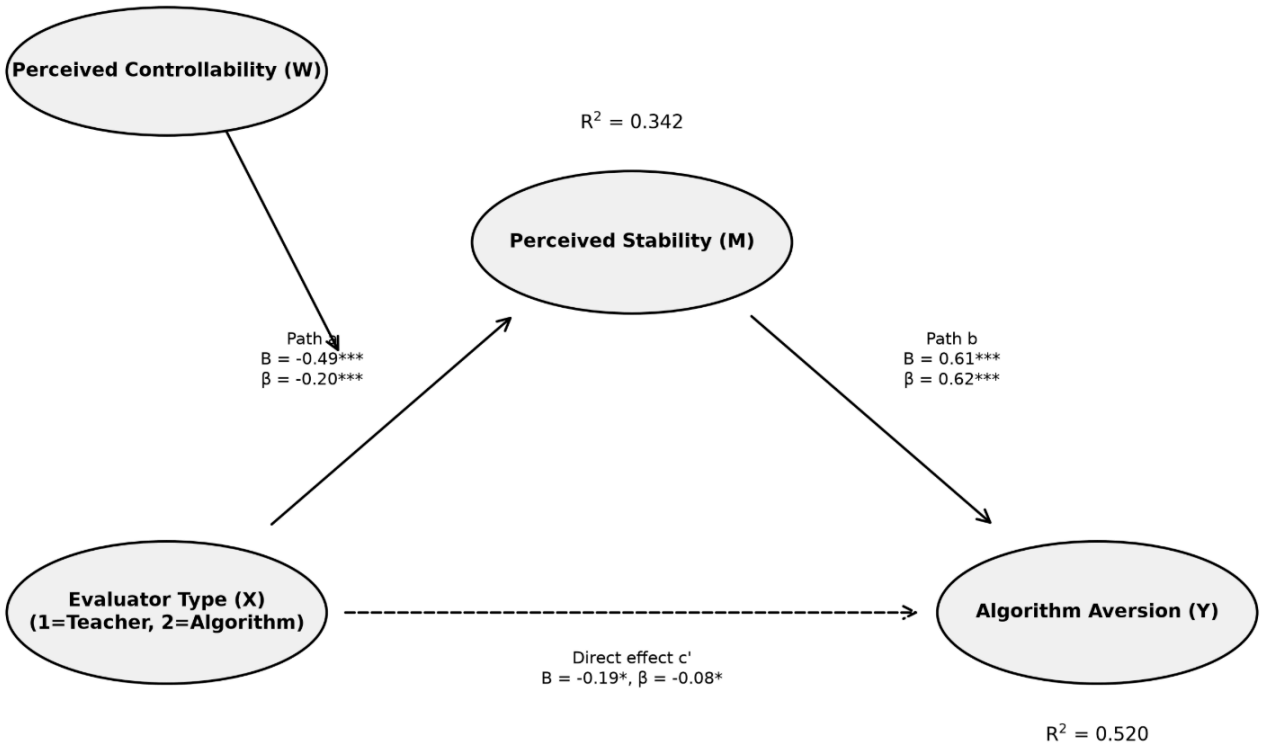

Figure 1

First-Stage Moderated Mediation Model of Algorithm Aversion in the Field of Academic Assessment (PROCESS Model 7)

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- file.xlsx