Evaluation Review
https://doi.org/10.1177/0193841X261451124

Journal Article

Transparency, Ethical Framing, and User Agency as Determinants of Trust in AI-Mediated Assessment: Informing the Design of Trustworthy Systems

Manuel B. Garcia [a,b,c] *

[a] College of Education, University of the Philippines Diliman, Quezon City, Philippines
[b] Educational Innovation and Technology Hub, FEU Institute of Technology, Manila, Philippines
[c] Graduate School of Education, Korea University, Seoul, South Korea

*** Correspondence:**
Manuel B. Garcia, University of the Philippines Diliman and FEU Institute of Technology.
mbgarcia@feutech.edu.ph

How to cite this article:
Garcia, M. B. (2026). Transparency, Ethical Framing, and User Agency as Determinants of Trust in AI-Mediated Assessment: Informing the Design of Trustworthy Systems. *Evaluation Review*. https://doi.org/10.1177/0193841X261451124.

Article History:
Received: 25 May 2025
Revised: 17 Nov 2025
Accepted: 4 Apr 2026
Published: 9 May 2026

Abstract:

As artificial intelligence (AI) systems assume greater responsibility in educational assessment, questions surrounding fairness, transparency, and trust have become central to their ethical and pedagogical legitimacy. Yet little empirical work has examined how specific design features shape students' trust in AI-driven assessment, particularly in contexts where algorithmic decisions carry meaningful academic consequences. This study examines how transparency, ethical framing, and user agency influence students' trust in an AI-based assessment platform. Using a 2×2×2 between-subjects experimental design with 240 undergraduate participants, the study isolates the main and interaction effects of these variables on trust, perceived fairness, perceived control, and adoption intention. Findings indicate that transparency is the most influential predictor of trust, while user agency functions as a compensatory mechanism in low-transparency conditions. Ethical framing, although theoretically salient, showed limited impact once users interacted with the system directly and shifted their attention toward the more concrete procedural cues embedded in the interface. A significant interaction between transparency and agency underscores the importance of aligning epistemic clarity with procedural control to foster behavioral commitment. These results support a multidimensional model of trust that incorporates emotional security, procedural justice, and behavioral intent. Overall, the study underscores that trust in AI assessment is not a byproduct of system accuracy alone but a reflection of students' perceived legitimacy of the evaluative process.

Keywords:

Educational Assessment, AI-Mediated Assessment, Trust in AI, Responsible AI Design, Digital Assessment, AI in Education

INTRODUCTION

Trust is a cognitive-affective stance of accepting vulnerability to a system based on expectations of its reliability, integrity, and normative alignment. In artificial intelligence (AI) research, trust is regarded as foundational element of effective human–AI interaction (Durán & Pozzi, 2025). As AI technologies become increasingly embedded in decision-making across domains, the need to understand and design for trust has become more urgent (Henrique & Santos, 2024). Despite a surge of empirical interest in trust in AI, the field remains conceptually fragmented, methodologically inconsistent, and disproportionately dominated by techno-centric perspectives. A recent twenty-four-year bibliometric review of empirical research on trust in AI by Benk et al. (2025) revealed that most studies remain exploratory, often lack theoretical grounding, and fail to integrate context-specific models that reflect the diversity of AI applications. In education, trust is frequently invoked as a prerequisite for successful AI adoption (e.g., Shata & Hartley, 2025), yet few studies investigate how domain-specific features shape trust formation. Recent work by Anagnostopoulou et al. (2024) has emphasized the importance of transparency, human oversight, and ethical safeguards in fostering trust in AI-driven learning environments, while Nazaretsky et al. (2025) has underscored the need for multidimensional trust measures tailored to educational settings. As trust in AI continues to be examined through broad conceptual lenses, there is a growing need to investigate how specific design features influence trust in educational settings where system outputs directly impact learners' experiences.

Among the many challenges introduced by the integration of AI in education (AIED), assessment has emerged as one of the most consequential and epistemically fraught domains (Luo, 2024; Stanoyevitch, 2024; Wachira et al., 2025). The increasing availability of AI tools capable of producing humanlike outputs has called into question the construct validity and epistemic integrity of conventional assessment practices. Garcia et al. (2025) noted that these systems can now generate essays, solve complex problems, and produce individualized responses that are often indistinguishable from student-authored work. Consequently, scholars have argued that traditional assessment paradigms are increasingly susceptible to appropriation by algorithmic agents, which can obscure the authentic demonstration of learner competence (Izquierdo-Álvarez & Jimeno-Postigo, 2025; Stanoyevitch, 2024; Swiecki et al., 2022). Parallel concerns have been raised regarding the algorithmic opacity of AI scoring systems, the attenuation of educator judgment, and the erosion of assessment's formative and dialogic functions in supporting student learning (Bozkurt et al., 2024; Corbin et al., 2025; Hasanah et al., 2025; Xiao et al., 2025). As AI systems assume a more constitutive role in determining what is assessed and how, urgent questions surface about the normative assumptions embedded in their design, the legitimacy of machinic evaluation, and the implications of these shifts for student trust and institutional accountability. These tensions point to a critical need for more empirical research that interrogates the sociotechnical consequences of AI-enabled assessment.

BACKGROUND OF THE STUDY

Transparency as a Design Imperative in Educational AI

Transparency is widely regarded as a cornerstone of trustworthy AI, particularly in contexts where users are directly impacted by algorithmic decisions (Buijsman, 2024; Felzmann et al., 2020; Radanliev, 2025). In educational settings, transparency encompasses not only the visibility of system processes but also the intelligibility of those processes to non-expert users (e.g., students and educators). Scholars argue that opaque AI systems undermine user confidence, compromise accountability, and exacerbate perceptions of arbitrariness in grading or feedback mechanisms (Balasubramaniam et al., 2023; Memarian & Doleck, 2023). While technical literature often frames transparency through explainability metrics and algorithmic interpretability, education researchers have emphasized its normative dimension by linking it to fairness, student autonomy, and pedagogical alignment (Chan, 2023; Memarian & Doleck, 2023). Yet despite broad acknowledgment of its importance, there remains limited empirical research on how specific transparency cues (e.g., detailed rubrics, scoring rationales, or visual feedback) shape students' trust in AI-mediated assessment systems. Given the increasing reliance on AI for formative and summative evaluation, transparency must be studied not merely as a technical attribute but as a design feature with psychological and ethical consequences. As assessment systems become more algorithmically mediated, this absence invites closer attention to the role transparency might play in shaping trust and interpretive legitimacy.

Ethical Framing as a Psychological Cue in Student-AI Interactions

As algorithmic systems increasingly mediate educational experiences, how these technologies are introduced and framed to users has emerged as a consequential ethical and psychological variable (Archambault et al., 2024; Fu & Weng, 2024; Yan & Liu, 2024). Ethical framing refers to the rhetorical emphasis placed on the benefits or risks of AI, often communicated through onboarding narratives, interface prompts, or institutional messaging. In educational settings, this framing functions not only as an ethical signal but also as a cognitive primer that may shape students' receptivity, scrutiny, and perceived legitimacy of AI-based decisions (Wiese et al., 2025). Studies have shown that the way AI is introduced can activate different mental models of the system, such as viewing it as a helpful assistant versus a surveillance instrument, with downstream effects on trust, acceptance, and critical engagement (Fu & Weng, 2024; Nguyen et al., 2023). Despite its salience, studies rarely isolate ethical framing as an experimental factor and often treats trust as an outcome of system performance rather than a product of sociotechnical context. This omission is especially problematic in educational assessment, where algorithmic authority intersects with fairness norms and learner vulnerability (Kamali et al., 2024). Given its role in shaping perceptions before interaction even begins, ethical framing warrants deeper investigation as a modifiable design feature capable of influencing trust dynamics in AI-supported learning environments.

User Agency as a Counterbalance to Algorithmic Authority

The growing integration of AI into educational infrastructures has sparked renewed concern over the erosion of human agency in decision-making systems (Adarkwah et al., 2025; Ouyang & Jiao, 2021). As AI begins to adjudicate not just what students learn, but how their performance is interpreted and acted upon, questions of control, contestability, and autonomy become increasingly central. User agency (sometimes referred to as human agency) refers to the capacity of individuals to influence, override, or challenge algorithmic outcomes, particularly in contexts where those outcomes carry evaluative or developmental consequences. In assessment, where the stakes are pedagogical as well as psychological, the absence of user agency may lead to feelings of disempowerment, procedural opacity, and diminished trust in system outputs (Fanni et al., 2023; Krakowski, 2025). Studies have shown that when users lack opportunities to interrogate or appeal algorithmic judgments, they are more likely to disengage or interpret those judgments as illegitimate (Garcia et al., 2025). Conversely, systems that embed mechanisms for human review or selective override have been associated with heightened perceptions of fairness, accountability, and trust. Despite its theoretical salience, user agency is rarely treated as a discrete design feature in empirical work on AI in education. Its potential to shape users' psychological orientation toward algorithmic assessment, particularly in environments where decisions feel automated and irreversible, merits closer scrutiny.

Synthesis, Research Gap, and Study Objectives

Transparency, ethical framing, and user agency have each been identified as salient design features that shape how individuals make sense of and respond to AI systems. Transparency enhances perceived fairness and interpretability, ethical framing shapes users' predispositions toward trust or skepticism, and agency reinforces a sense of control, contestability, and procedural legitimacy. While each construct has been studied in relation to trust, the existing literature remains fragmented, with most research examining these factors in isolation and often outside of educational or high-stakes evaluative contexts. Yet in educational settings, where algorithmic judgments intersect with learners' epistemic vulnerability and institutional accountability, the dynamics of trust may differ substantially from those observed in general human–AI interaction research. This siloed approach neglects the possibility that trust is not merely a response to individual system features, but a composite judgment shaped by the interaction of multiple, co-occurring design cues. Moreover, few studies have empirically tested how these cues function within the specific context of AI-supported assessment, where system judgments carry institutional weight and pedagogical significance. To address this gap, the study investigates the independent and interactive effects of transparency, ethical framing, and user agency on students' trust in an AI-mediated assessment platform. By experimentally isolating and combining these features, the study provides empirical insight into how interface-level decisions influence trust formation in educational AI, particularly in environments where perceptions of fairness, control, and credibility are central to learner experience.

METHODS

Research Design

This study employed a between-subjects experimental design with a 2 (Transparency: High vs. Low) × 2 (Ethical Framing: Risk vs. Benefit) × 2 (User Agency: Present vs. Absent) factorial structure. The approach was selected to enable causal inference regarding the effects of key features on user trust in AI-based assessment systems. By systematically manipulating the independent variables and randomly assigning participants to one of eight experimental conditions, the study ensured control over potential confounding variables and facilitated the analysis of both main effects and higher-order interaction effects. This factorial structure was particularly appropriate for investigating not only the isolated effects of each design dimension but also their potential combinatorial influence. In emerging sociotechnical systems where user perceptions are often shaped by interdependent cues, understanding such interaction effects is essential for informing responsible and trustworthy AI design. The primary dependent variable was trust in the AI system, assessed using validated psychometric instruments. Manipulation checks were administered to verify that participants accurately perceived the intended levels of transparency, ethical framing, and user agency. All procedures conformed to the research ethics protocols of the host institution and were conducted in accordance with the Declaration of Helsinki. Participation was voluntary, responses were anonymized, and participants were debriefed following the study session. No deception was used, and all participants were fully informed about the purpose and nature of the AI system they interacted with.

Setting and Participants

The study was conducted across three campuses located in different regions but affiliated under the same university system. Participants were selected from undergraduate students enrolled in writing-intensive courses such as *Writing for New Media*, *Scriptwriting and Storyboarding*, *Reading into Writing*, and *Business Writing and Communications*. These courses were chosen because their assessment formats aligned closely with the simulated AI-mediated essay grading environment used in the experiment. To qualify for participation, students had to be currently enrolled in one of the specified courses and demonstrate fluency in English. Individuals with professional experience in AI development (i.e., students from computing-related programs) were excluded to minimize potential bias from domain-specific knowledge. A power analysis conducted in G*Power (α = .05, *power* = .80, *medium effect size f* = 0.25) indicated a minimum required sample size of 128. To improve statistical robustness and allow for attrition, a total of 240 participants were recruited and randomly assigned to one of the eight experimental conditions (n = 30 per condition). The final sample comprised undergraduate students, with a mean age of 20.3 years (SD = 1.06). Gender representation included 57% female, 42% male, and 1% non-binary. All participants provided informed consent before participation.

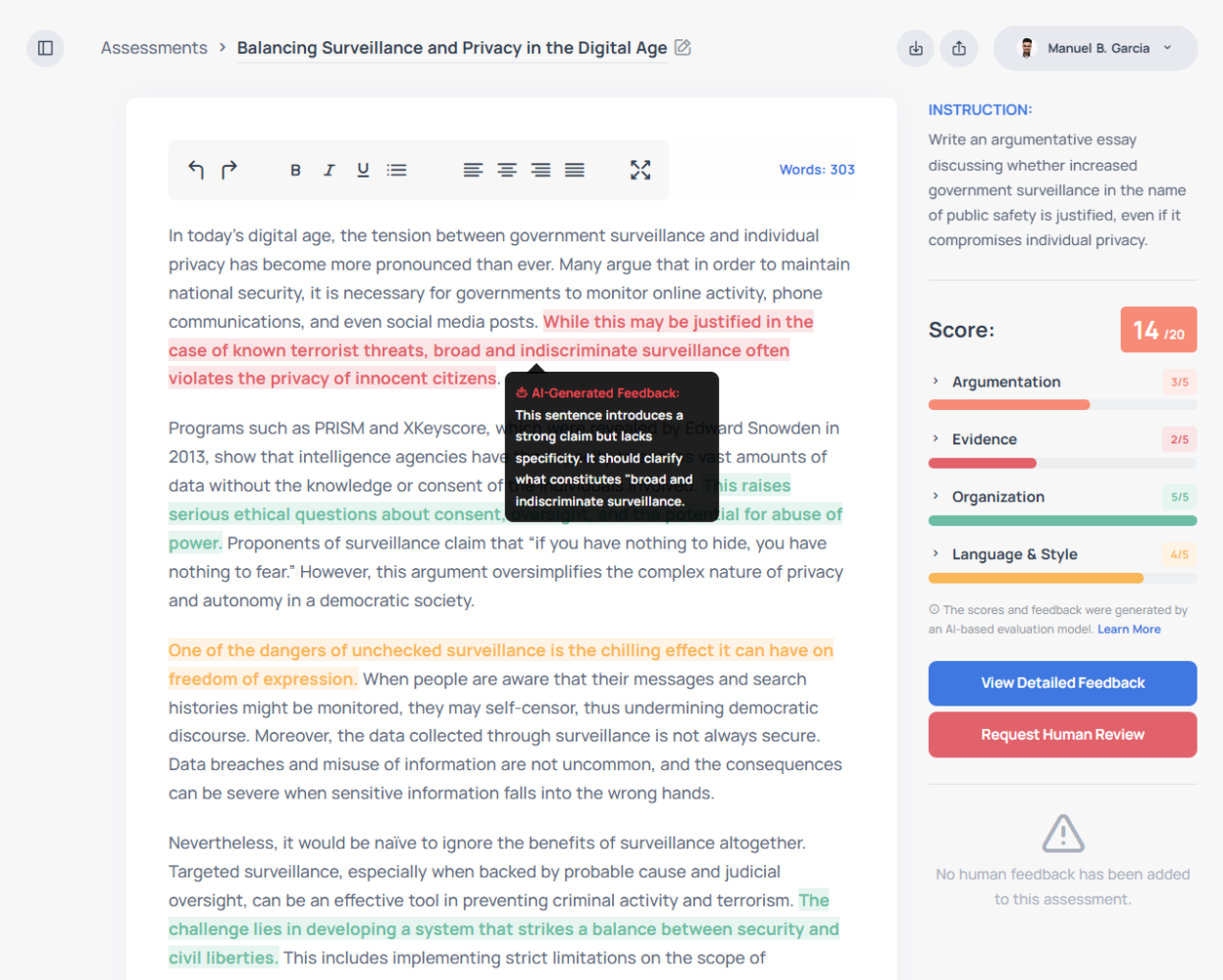

Figure 1. Simulated AI-mediated assessment interface used in the experimental platform

Experimental Procedures

The experiment was conducted using a custom-designed online simulation platform that emulated an AI-mediated essay grading environment (see Figure 1). The platform was powered by a backend that incorporated natural language generation components and pre-scripted decision logic to simulate how an AI system might evaluate student essays. The AI feature of the platform was powered by the ChatGPT API, which was used to generate natural language explanations simulating how an AI system might justify grading decisions. However, all scoring outputs were standardized and pre-scripted across conditions to maintain experimental control. Upon accessing the simulation, participants were introduced to the scenario in which an AI tool had graded their work. Each participant was shown a pre-written, anonymized argumentative essay on a contemporary social issue such as the ethical use of surveillance technologies or universal basic income. The essay topic was designed to be broadly relevant across disciplines but framed in a way that reflected the students' program area. Participants were instructed to assume the role of

a student receiving AI-generated feedback and a corresponding grade for the essay. The specific manipulations of transparency, ethical framing, and user agency are summarized in Table 1.

Table 1. Summary of Experimental Manipulations Embedded in the AI-Based Assessment Platform

Experimental Factor	Condition	Operationalization within the Simulation Platform	Purpose/Rationale
Transparency	High Transparency	Displayed a detailed scoring rubric showing how points were allocated across argument structure, evidence, organization, and style, with accompanying natural language explanations simulating AI reasoning.	To enhance system interpretability, support procedural fairness, and foster cognitive-based trust.
	Low Transparency	Displayed only a final grade without any rubric or justification.	To simulate opaque algorithmic systems and examine trust under conditions of limited explainability.
Ethical Framing	Benefit-Oriented	Onboarding message highlighted the AI system's advantages: fairness, efficiency, and scalability in assessment.	To prime positive expectations and institutional legitimacy of AI in education.
	Risk-Oriented	Onboarding message emphasized potential drawbacks: bias, misuse, and lack of accountability in algorithmic grading.	To prime critical awareness and ethical concern, prompting reflective skepticism toward AI systems.
User Agency	Agency Present	Included a clickable option allowing participants to request a human review of the AI-generated grade.	To simulate human oversight, increase perceived control, and test the effect of user agency on trust.
	Agency Absent	Informed participants that the AI grade was final and could not be contested.	To simulate fully autonomous decision-making systems and examine trust in the absence of recourse options.

Data Collection and Analysis

Data were collected through an integrated survey system on the same platform that hosted the simulation. The primary dependent variable was measured using a validated instrument developed by Nazaretsky et al. (2025) designed to assess student trust in AI-mediated educational technologies. The original instrument comprises 20 items across four subscales: perceived usefulness, perceived obstacles, perceived readiness, and perceived trust. For the purposes of this study, the original 5-point Likert scale was adapted to a 7-point format (1 = Strongly Disagree, 7 = Strongly Agree) to increase response sensitivity and align with the measurement scales used for other variables in the study. Internal consistency remained acceptable across all subscales, with Cronbach's alpha values ranging from .76 to .88. Supplementary variables included perceived control, perceived fairness, and intention to use AI in future assessments. These variables were included to explore how trust-related perceptions may extend to behavioral and fairness-related judgments, which are often implicated in students'

acceptance and interpretation of AI-assisted assessment systems. The same factorial model was applied to these secondary outcomes to examine whether the experimental manipulations extended beyond trust. Data cleaning involved excluding participants with incomplete responses and those who failed attention or manipulation checks. Descriptive statistics were first computed, followed by a three-way analysis of variance (ANOVA) to examine the main and interaction effects of transparency, ethical framing, and user agency on trust. Where significant interactions emerged, simple effects analyses and Tukey's post hoc comparisons were conducted. Effect sizes were reported using partial eta squared, and statistical significance was set at the .05 level.

RESULTS

Manipulation checks confirmed that participants accurately perceived the experimental conditions to which they were assigned. Across all groups, correct identification rates were high for transparency (92.4%), ethical framing (91.3%), and user agency (94.1%). A series of one-way ANOVAs revealed no significant differences in manipulation check accuracy across experimental groups: transparency, $F(1, 238) = 1.14$, $p = .287$; ethical framing, $F(1, 238) = 1.02$, $p = .314$; and agency, $F(1, 238) = 0.89$, $p = .346$. The high accuracy rates and absence of systematic variation across conditions indicate that the experimental manipulations were reliably implemented and perceived as intended. As shown in Table 2, mean trust scores varied systematically across conditions, with the highest trust observed in the high-transparency, benefit-framing, agency-present group ($M = 5.43$, $SD = 0.82$). In contrast, the lowest trust was reported in the low-transparency, risk-framing, agency-absent group ($M = 4.49$, $SD = 0.96$). These patterns support the perceptual salience and theoretical relevance of the manipulated variables.

Table 2. Mean Trust Scores by Experimental Condition

Condition	Transparency	Framing	Agency	Mean Trust	SD
1	High	Benefit	Present	5.43	0.82
2	High	Benefit	Absent	5.39	0.85
3	High	Risk	Present	5.12	0.88
4	High	Risk	Absent	4.81	0.91
5	Low	Benefit	Present	5.03	0.87
6	Low	Benefit	Absent	4.43	0.90
7	Low	Risk	Present	4.71	0.94
8	Low	Risk	Absent	4.49	0.96

Main Effects of Transparency, Ethical Framing, and User Agency

A three-way between-subjects ANOVA was conducted to examine the effects of transparency, ethical framing, and user agency on users' trust in the AI-based assessment system (see Table 3). The analysis revealed a significant main effect of transparency, $F(1, 232) = 22.18$, p

< .001, η^2_p = .087. Participants in the high-transparency condition, who received detailed scoring rubrics accompanied by natural language explanations, reported significantly higher trust (*M* = 5.19, *SD* = 0.88) compared to those in the low-transparency condition, who were shown only a final grade without justification (*M* = 4.66, *SD* = 0.92). Similarly, there was a significant main effect of user agency, *F*(1, 232) = 9.62, *p* = .002, η^2_p = .040. Participants who were given the ability to challenge the AI's decision reported significantly higher trust (*M* = 5.07, *SD* = 0.89) compared to those who were not given this option (*M* = 4.76, *SD* = 0.91). Contrary to expectations, the main effect of ethical framing was not statistically significant, *F*(1, 232) = 1.93, *p* = .166, η^2_p = .008. Participants who received a benefit-oriented message describing the advantages of AI (*M* = 5.06, *SD* = 0.90) did not report significantly different levels of trust compared to those exposed to a risk-oriented framing that emphasized potential harms (*M* = 4.85, *SD* = 0.91). These findings suggest that initial messaging, while potentially influential at a rhetorical level, may not be sufficient to shape trust once users engage directly with the system and observe its behavior. These results reinforce that procedural features (e.g., clarity and control) carry greater weight in trust formation than rhetorical or introductory cues.

Table 3. ANOVA Summary for Trust in AI-Based Assessment

Source	*F*	*p*	Partial η^2
Transparency	22.18	< .001	.087
Ethical Framing	1.93	.166	.008
User Agency	9.62	.002	.040
Transparency × Framing	0.87	.353	.004
Transparency × Agency	6.11	.014	.026
Framing × Agency	1.04	.308	.004
Transparency × Framing × Agency	3.91	.055	.017

Interaction Effects Among Experimental Factors

Table 3 also illustrates complex interaction patterns among the experimental variables. The analysis revealed a significant interaction between transparency and user agency, *F*(1, 232) = 6.11, *p* = .014, η^2_p = .026. Post hoc comparisons indicated that within the low-transparency condition, participants who had the option to request a human review reported significantly higher trust (*M* = 4.87, *SD* = 0.87) than those who were not given this option (*M* = 4.46, *SD* = 0.91), *p* = .008. In the high-transparency condition, trust levels remained relatively high regardless of agency, with no significant difference between groups (*M* = 5.28 vs. *M* = 5.10), *p* = .210. These findings suggest that user agency may serve as an essential compensatory mechanism in contexts where transparency is limited. When the system does not provide sufficient explanation, the ability to challenge the AI's decision appears to help restore trust. However, when transparency is already present, additional control over the outcome does not substantially enhance trust. Meanwhile, no significant two-way interactions were observed

between transparency and ethical framing or between ethical framing and user agency. A marginally significant three-way interaction among transparency, ethical framing, and agency was also identified, $F(1, 232) = 3.91$, $p = .055$, $\eta^2_p = .017$. Exploratory comparisons showed that in the condition combining low transparency and risk-oriented framing, participants reported higher trust when agency was present ($M = 4.71$) compared to when it was absent ($M = 4.49$). In contrast, under high transparency and benefit-oriented framing, agency had little impact on trust ($M = 5.43$ vs. $M = 5.39$). This finding suggests that user agency may be particularly effective in enhancing trust when both procedural clarity and ethical assurance are lacking. In such conditions, agency may not only provide a sense of control but also offer psychological reassurance against potential system fallibility.

Secondary Outcomes: Fairness, Control, and Adoption Intention

Table 4 reveals clear and consistent patterns across the secondary outcomes. It shows that the combined presence of transparency and user agency produces the most favorable user perceptions. Participants in the high-transparency and agency-present condition consistently reported the highest ratings. In this group, the mean score for perceived fairness was 5.31 (SD = 0.81), perceived control was 5.26 (SD = 0.79), and intention to use AI was 5.41 (SD = 0.83). In contrast, participants in the low-transparency and agency-absent condition reported the lowest scores across all three outcomes, with a mean fairness rating of 4.33 (SD = 0.90), control rating of 4.18 (SD = 0.92), and adoption intention of 4.38 (SD = 0.89). Intermediate values were observed in the mixed conditions, with ratings generally reflecting the additive influence of transparency and agency. For example, when transparency was high, but agency was absent, perceived fairness averaged 5.07 (SD = 0.88), control was 4.89 (SD = 0.84), and adoption intention was 5.12 (SD = 0.90). When transparency was low, but agency was present, fairness was rated at 4.71 (SD = 0.86), control at 4.63 (SD = 0.91), and adoption at 4.90 (SD = 0.87). These descriptive patterns further support the conclusion that the combination of high transparency and user agency leads to the most favorable perceptions of AI systems in educational assessment.

Table 4. ANOVA Summary for Secondary Outcomes

Dependent Variable	Factor	F	p	Partial η²
Perceived Fairness	Transparency	17.92	< .001	.072
	Agency	2.31	.129	.010
	Framing	0.98	.323	.004
	Transparency × Agency	1.21	.272	.005
	Transparency × Framing	0.88	.350	.004
	Agency × Framing	0.73	.393	.003
	Transparency × Agency × Framing	0.91	.341	.004
Perceived Control	Agency	11.03	.001	.045
	Transparency	1.56	.213	.007

	Framing	0.72	.397	.003
	Transparency × Agency	1.14	.287	.005
	Transparency × Framing	0.69	.408	.003
	Agency × Framing	0.81	.369	.003
	Transparency × Agency × Framing	1.06	.304	.005
Intention to Use AI	Transparency	2.84	.093	.012
	Agency	3.01	.084	.013
	Framing	0.89	.347	.004
	Transparency × Agency	8.45	.004	.035
	Transparency × Framing	1.34	.249	.006
	Agency × Framing	0.97	.326	.004
	Transparency × Agency × Framing	0.83	.362	.004

Inferential analysis supported these trends. Perceived fairness of the AI system was significantly influenced by transparency, $F(1, 232) = 17.92$, $p < .001$, $\eta^2_p = .072$. Participants in the high-transparency condition rated the system as significantly fairer than those in the low-transparency group. Perceived control was most strongly influenced by user agency, $F(1, 232) = 11.03$, $p = .001$, $\eta^2_p = .045$, with participants who were given the option to challenge the output reporting a greater sense of control than those in the agency-absent condition. For intention to use AI-based assessment systems, the main effects of transparency and agency were not statistically significant, although both trended in the expected direction, $F(1, 232) = 2.84$, $p = .093$ and $F(1, 232) = 3.01$, $p = .084$, respectively. Ethical framing also showed no significant effect, $F(1, 232) = 0.89$, $p = .347$. However, a significant interaction between transparency and agency was observed, $F(1, 232) = 8.45$, $p = .004$, $\eta^2_p = .035$. The highest adoption intention was reported when both transparency and agency were present ($M = 5.41$), while the lowest was recorded when both were absent ($M = 4.38$). This interaction suggests that these two design features may operate synergistically to increase users' willingness to adopt AI systems in educational contexts.

DISCUSSION

Trust is a foundational construct in the responsible integration of AI into education, particularly in domains where automated systems exercise evaluative authority. As AI tools are increasingly deployed to support or supplant human judgment in assessment, questions surrounding procedural fairness, system transparency, and user autonomy have become design imperatives. Yet despite widespread consensus on the importance of fostering trust in educational AI, empirical investigations remain limited and often constrained by reductive models that treat trust as a static outcome of isolated system features. This study addressed that gap by examining how three theoretically significant and interface-level factors independently and interactively

shape students' trust in an AI-mediated assessment system. Positioning transparency, ethical framing, and user agency as modifiable components of system design advances an interactionist view of trust that accounts for both the discrete influence of individual features and the emergent effects of their configuration. Unlike prior work that often conceptualizes trust in monolithic terms, this approach recognizes trust as a context-sensitive response shaped by users' interpretations of procedural clarity, value alignment, and participatory affordances. In AI-based assessment, understanding how these features intersect is essential for developing systems that are not only technically robust but also epistemically credible and ethically defensible.

Synergistic Effects of Transparency and Agency on Adoption Intention

At the core of these findings is the interaction between transparency and agency, which reveals how their combination drives students' behavioral intent. This interaction underscores that adoption decisions in AI-based assessment are contingent on the alignment between system intelligibility and perceived volitional control. When users encounter systems that are both procedurally transparent and offer meaningful avenues for human intervention, their intention to engage with the system increases substantially (Kelly et al., 2023; Radanliev, 2025). This finding reflects the reciprocal affordance hypothesis, where the presence of one design affordance (e.g., transparency) amplifies the perceived value of another (e.g., agency). Behavioral intention in such contexts is shaped not only by what the system does, but by how well its affordances match users' normative expectations for control, fairness, and interpretability. This pattern is consistent with broader findings in AI adoption research, where task-technology fit and trust have been shown to outperform traditional usability constructs in predicting engagement (Garcia, 2025). When users perceive a system as both understandable and influenceable, they develop a stronger sense of cognitive alignment and procedural justice. Wanner et al. (2022) similarly argue that adoption is most likely when systems afford not just passive clarity but also active involvement where users can make sense of outcomes and exert corrective influence. In this light, the synergy between transparency and agency fosters a value-aligned commitment to the system. Rather than being driven by compliance or novelty, adoption becomes a function of perceived legitimacy, interpretive coherence, and user empowerment (Martínez-Requejo et al., 2025). These findings reinforce the view that successful AI integration in education depends not only on system performance but on the intentional orchestration of affordances that uphold users' epistemic agency and ethical expectations throughout their interaction with the technology.

Transparency as a Primary Driver of Trust and Fairness

Given its pronounced and consistent effects across outcomes, transparency occupies a central position in shaping students' evaluations of AI-mediated assessment. This finding affirms that transparency is not merely a design convenience but a foundational precondition for trust and fairness in AI-mediated assessment. Participants exposed to high-transparency conditions reported significantly greater trust and fairness. These results underscore the role of transparency as both an epistemic and normative affordance. As posited by Buijsman (2024), value-based transparency provides legitimacy by revealing the moral reasoning embedded in

system design. This approach enables users to interpret outcomes as procedurally just rather than arbitrary. Supporting this claim is the "Transparency by Design" framework (Felzmann et al., 2020), which conceptualizes transparency as a relational mechanism that enables users to make sense of and act upon algorithmic judgments. In educational contexts, such framing is particularly salient, as it enables learners to form coherent mental models of how and why decisions are made. This alignment between system logic and learner expectations enhances perceptions of procedural justice, especially in high-stakes environments where evaluation outcomes carry formative or summative weight. Importantly, the form of transparency operationalized in this study maps onto Memarian and Doleck's (2023) distinction between descriptive and technical transparency, with the former proving especially effective in shaping trust perceptions among non-expert users. Consistent with Larsson and Heintz's (2020) view of transparency as a conceptual metaphor for system knowability, the finding that transparency significantly influences both trust and fairness suggests it functions as a cognitive scaffold for interpretive legitimacy. Collectively, these findings demonstrate that transparency, when embedded as a communicative and contextual design function rather than as a retrospective add-on, enhances interpretability, affirms procedural integrity, and anchors students' trust in AI-based assessments.

Agency as a Psychological Buffer in Low-Transparency Environments

Agency functioned less as a direct driver of trust and more as a feature whose impact depended on the degree of transparency provided. Accordingly, while transparency emerged as the primary driver of trust, the findings also point to the compensatory role of user agency in offsetting the adverse effects of low-transparency environments. In conditions where algorithmic decision-making remained opaque, the mere presence of an opt-out or human-review option significantly enhanced perceptions of fairness and procedural legitimacy. This result aligns with the dual-perspective theory of agency in human–AI interaction (Legaspi et al., 2024), which argues that preserving a user's first-person sense of agency is essential for maintaining engagement and perceived authorship over outcomes. When AI systems dominate the decision space without affording users corrective recourse, users' sense of control diminishes (Ouyang & Jiao, 2021). This loss of control is frequently associated with disengagement and reduced accountability. In contrast, agency-restoring mechanisms (e.g., the ability to request redress or contest automated outcomes) serve as psychological buffers that reinstate a minimal but meaningful form of volitional influence. This interpretation is consistent with the concept of active human agency (Adarkwah et al., 2025; Fanni et al., 2023), wherein mechanisms for contestability and redress re-anchor users in the interpretive loop. From a systems design standpoint, this aspect substantiates the argument that agency is a cognitive and affective scaffold that reinforces human–system co-regulation (Krakowski, 2025). By operationalizing even lightweight agency affordances, designers can enhance user autonomy without undermining algorithmic efficiency. In contexts where full transparency is unattainable, embedding contestability mechanisms becomes a necessary design strategy. Overall, the findings suggest that user agency serves as a psychological counterweight to the opacity inherent in many AI systems when supported through structured opportunities for intervention.

Ethical Framing and the Limits of Initial Impressions

Perhaps unexpectedly, the results revealed no significant main effect of ethical framing on trust. This finding invites a closer examination of the limitations of initial moral positioning in AI-mediated assessment. While ethical onboarding messages may shape first impressions, their influence appears to dissipate when users engage directly with system features that more tangibly convey procedural justice. Fu and Weng (2024) emphasize that responsible human-centered AI cannot rely solely on front-loaded ethical cues but must be reinforced through meaningful system interactions that embody fairness, autonomy, and intelligibility across time. This interpretation aligns with Holmes et al. (2022), who argue that doing things ethically, not merely doing ethical things, requires ethical design to permeate every layer of AIED systems, from their computational logic to their pedagogical goals. In educational contexts, the perceived ethical integrity of a system is shaped less by rhetorical declarations and more by operational coherence between ethical intent and actual function. Dignum (2019) similarly distinguishes between ethics by intention and ethics by design and asserted that credibility arises not from professed values but from their procedural realization. Collectively, these perspectives underscore the interpretive fragility of ethical framing. While it may serve as a priming device, its influence remains marginal without reinforcement through tangible procedural fairness and opportunities for user agency.

Beyond these theoretical explanations, several contextual and methodological considerations may help account for why ethical framing did not exert a measurable effect in this study. First, the brevity and placement of the framing manipulation may have limited its salience relative to the more consequential system interactions that followed. Participants were required to engage directly with the assessment interface and its outputs, which may have overshadowed the initial framing message and shifted their attention toward operational qualities. As reinforced in a meta-review (Bond et al., 2024), ethical considerations that are treated as peripheral rather than embedded into the design and enactment of AI systems exert limited influence on how users interpret, evaluate, and engage with AIEd. Second, the low-stakes and short-term nature of the simulated assessment may have attenuated sensitivity to value-based positioning, as users tend to prioritize functional clarity over moral assurances when evaluative consequences are minimal. Third, cohort-specific familiarity with digital platforms may have diminished the perceived relevance of ethical messaging, particularly when such statements are routinely encountered in everyday technologies and thus carry reduced interpretive weight. Finally, ethical framing may exert its influence cumulatively rather than instantaneously. Nguyen et al. (2023) note that trust is built through iterative interpretive encounters, where system transparency and accountability are observed in action rather than assumed from framing. These considerations suggest that the null effect may reflect not an absence of influence in principle but the limits of brief, standalone framing cues within short-term experimental interactions.

Trust as a Multidimensional Judgment in AI Assessment

As evidenced in AI-mediated assessment contexts, trust emerges from a constellation of interrelated judgments encompassing affective comfort with automation, perceptions of

procedural fairness, and behavioral willingness to endorse or continue using the system. Characterizing trust in this way reflects a shift away from reductive definitions that treat it as mere acceptance or passive system compliance. Trust in AI-mediated assessment is not a unitary response but an integrated judgment shaped by students' interpretive evaluations of clarity, fairness, and autonomy throughout their interaction with the system. Although the study did not employ a predefined model of trust, the patterns that emerged reveal how the three manipulated factors collectively guided students' sensemaking processes. In this light, trust develops through the alignment of these experiential cues rather than from any isolated feature. Clarity about evaluative logic, opportunities for corrective intervention, and coherence between ethical intent and system behavior jointly informed whether students regarded the AI system as credible and worthy of reliance. Consistent with Nazaretsky et al. (2025), trust in educational AI is formed through overlapping cognitive evaluations and social-psychological responses, but in this study such evaluations were shaped directly by the experimentally manipulated features rather than by predetermined trust dimensions. The findings suggest that no single system feature fully secures trust unless it simultaneously addresses users' emotional comfort, procedural expectations, and behavioral readiness. Thus, designing for trust entails not just technical robustness or ethical declarations but the orchestration of experiential cues that affirm fairness, empathy, and user empowerment at every stage of interaction (Memarian & Doleck, 2023; Radanliev, 2025). In this sense, trust becomes a cumulative judgment that evolves over time and is essential to the legitimacy and long-term sustainability of AI-based assessment in education.

System-Level Design Principles for Trustworthy AI-Mediated Evaluation

Translating these findings into practice requires attention to the system-level design principles that enable trust to emerge within AI-mediated assessment. Building on the converging evidence from this study, a pragmatic agenda for designing AI-mediated assessment systems must foreground trust as an evolving judgment shaped by interactional fidelity, cognitive fit, and procedural integrity. Trust in educational AI does not arise from isolated technical features but from the deliberate orchestration of user-centered design elements that resonate with students' normative expectations for fairness, autonomy, and intelligibility. As Afroogh et al. (2024) emphasize, the cultivation of trust requires design strategies that integrate both technical robustness and axiological dimensions (e.g., transparency, accountability, and ethical alignment).

What emerges from the findings of this study is a triad of design commitments that function as the scaffolding for a trustworthy AI evaluation ecosystem. First, transparency should occur in advance to help users grasp how the system works before receiving outcomes instead of being explained only afterward. Proactive transparency not only clarifies evaluative logic but also establishes the epistemic grounding users need to interpret decisions as legitimate. Second, agency must be preserved through mechanisms that allow users to contest or override system decisions and provide a sense of psychological security despite algorithmic opacity. Such avenues for redress ensure that users remain active participants in the evaluative process rather than passive recipients of automated judgment. Third, procedural structures that provide evaluative

criteria, reasons for scoring, and avenues for appeal are necessary for making AI decisions clear and adjustable. These structures convert assessment from a static output into an interpretable and revisable dialogue between learner and system. Collectively, this approach aligns with the tripartite model of ethics by, in, and for design (Dignum et al., 2018), which underscores the need for ethical reasoning to be embedded not only within the system's computational logic but also across its development process and societal purpose. Designing for trust is not an afterthought but an ethical and architectural imperative for ensuring that AI-mediated evaluations are not only technically reliable but also pedagogically legitimate and socially defensible.

Learner-Level Implications for Fairness, Autonomy, and Student Experience

At the experiential level, the findings also invite a reconsideration of how AI-mediated assessment shapes students' perceptions of fairness, autonomy, and their overall engagement with evaluation. The integration of AI into educational assessment calls for a deeper reconsideration of what constitutes fairness, autonomy, and a meaningful student experience (Garcia et al., 2025). In AI-mediated evaluative environments, fairness cannot be conflated with procedural consistency alone. It must also encompass epistemic transparency and affordances for interpretive engagement, whereby students can interrogate the logic underlying their evaluations. Consider a scenario in which a student receives an anomalously low score on an AI-graded essay. If the system offers only a numeric score with no accompanying rationale or path for contesting the result, the student may not only question the fairness of the grade but also disengage from the learning process entirely. Conversely, when the system provides explanatory feedback and facilitates dialogic interaction, the assessment becomes an opportunity for learning rather than a site of evaluative alienation. Autonomy, within this framework, extends beyond logistical self-determination and entails a learner's participatory agency in the interpretive dimensions of assessment. It requires that students co-construct the evaluative process through mechanisms of feedback, contestability, and revision. AI systems should be designed to scaffold this form of autonomy by embedding procedural transparency, enabling auditability, and preserving instructional accountability. As Bond et al. (2024) emphasize, ethical frameworks must be embedded into system interactions if they are to shape student experience in ways that are pedagogically responsible. When students are structurally excluded from the interpretive arc of evaluation, the perceived legitimacy of the system deteriorates. However, when assessment architectures are designed to be transparent, contestable, and responsive, they foster procedural trust and reinforce the learner's sense of epistemic agency and educational belonging.

Limitations and Future Directions

While this study contributes to a growing body of work on trust in AI-mediated assessment, several limitations delimit the scope of its conclusions and point to productive avenues for future inquiry. First, the low-stakes assessment environment may have attenuated the influence of design features that operate more powerfully when evaluative consequences carry meaningful academic or affective weight. Future research should examine whether the patterns observed here replicate in authentic assessment contexts where students experience real

performance contingencies and heightened epistemic vulnerability. Second, although the ethical framing manipulation was theoretically grounded, its brevity and placement may have constrained its salience relative to the more consequential system interactions that followed. The null effect observed for ethical framing therefore warrants more nuanced investigation. Subsequent studies might explore alternative framings (e.g., narrative, dialogic, or context-sensitive), varied temporal placements (e.g., interleaved with system use), or repeated exposure across multiple interactions. Longitudinal designs could also shed light on whether ethical positioning exerts cumulative influence over time. Third, the study utilized a controlled interface with fixed operational parameters, which allowed for precise manipulation of the target features but necessarily narrowed ecological variability. Real-world AI assessment systems exhibit more complex, adaptive, and sometimes opaque behaviors. Future research should therefore examine how transparency, agency, and ethical cues operate in more dynamic environments, including platforms that employ continuous updating, richer feedback modalities, or adaptive scoring pipelines. Finally, while this study investigated interface-level design features, trust in AI-based assessment is also shaped by broader sociotechnical considerations (e.g., institutional policies, classroom practices, and instructor mediation). Future work could adopt a multilevel approach that integrates system-level affordances with pedagogical, organizational, and sociocultural dynamics to build a more comprehensive theory of trustworthy AI-mediated evaluation.

CONCLUSION

In the age of algorithmic instruction and machinic evaluation, educational assessment is undergoing a paradigmatic shift that demands new epistemic and ethical frameworks. As AI systems assume greater authority in grading and feedback, the criteria for effective assessment must expand beyond accuracy and efficiency to include legitimacy, transparency, and learner trust. This study advances the understanding of trust in AI-mediated evaluation as a multidimensional construct encompassing affective alignment, procedural fairness, and behavioral engagement. Rather than emerging from isolated design features, trust is formed through the cumulative interaction of system affordances that shape students' perceptions of interpretability and agency. The empirical findings demonstrate that transparency is the most robust determinant of trust, while user agency operates as a compensatory mechanism in contexts of limited explainability. Ethical framing, although salient at the rhetorical level, proves insufficient in the absence of deeper procedural integration. The interaction between transparency and agency reveals that trust and adoption intention are contingent on both epistemic clarity and perceived control. These insights support a model of assessment design grounded in pedagogical integrity and ethical responsibility. For developers, this entails embedding transparency into the system's operational logic, operationalizing user agency through contestability features, and integrating ethics not as peripheral messaging but as an intrinsic design property. For educators and institutional leaders, the sustainability of AI-assisted assessment will depend on whether students perceive these systems as not only functionally competent but also as procedurally just and pedagogically inclusive. As students increasingly

encounter automated decisions that affect their academic progression, institutions must act decisively to ensure that AI systems are designed not merely to assess learning, but to support it in ways that are transparent, participatory, and human-centered.

REFERENCES

Adarkwah, M. A., Amponsah, S., Huang, R., & Thomas, M. (2025). *Artificial Intelligence and Human Agency in Education: The Nexus Between AI and Human Agency in Educational Contexts*. Springer Singapore. https://doi.org/10.1007/978-981-96-7937-9

Afroogh, S., Akbari, A., Malone, E., Kargar, M., & Alambeigi, H. (2024). Trust in AI: Progress, Challenges, and Future Directions. *Humanities and Social Sciences Communications*, *11*(1), 1-30. https://doi.org/10.1057/s41599-024-04044-8

Anagnostopoulou, E., Grammatikos, N. A., Apostolou, D., & Mentzas, G. (2024). Trustworthy AI in education: A Roadmap for Ethical and Effective Implementation. *Proceedings of the 13th Hellenic Conference on Artificial Intelligence*. https://doi.org/10.1145/3688671.3688781

Archambault, S. G., Ramachandran, S., Acosta, E., & Fu, S. (2024). Ethical Dimensions of Algorithmic Literacy for College Students: Case Studies and Cross-Disciplinary Connections. *The Journal of Academic Librarianship*, *50*(3), 1-20. https://doi.org/10.1016/j.acalib.2024.102865

Balasubramaniam, N., Kauppinen, M., Rannisto, A., Hiekkanen, K., & Kujala, S. (2023). Transparency and Explainability of AI Systems: From Ethical Guidelines to Requirements. *Information and Software Technology*, *159*, 1-15. https://doi.org/10.1016/j.infsof.2023.107197

Benk, M., Kerstan, S., von Wangenheim, F., & Ferrario, A. (2025). Twenty-Four Years of Empirical Research on Trust in AI: A Bibliometric Review of Trends, Overlooked Issues, and Future Directions. *AI & SOCIETY*, *40*(4), 2083-2106. https://doi.org/10.1007/s00146-024-02059-y

Bond, M., Khosravi, H., De Laat, M., Bergdahl, N., Negrea, V., Oxley, E.,...Siemens, G. (2024). A Meta Systematic Review of Artificial Intelligence in Higher Education: A Call for Increased Ethics, Collaboration, and Rigour. *International Journal of Educational Technology in Higher Education*, *21*(1), 1-41. https://doi.org/10.1186/s41239-023-00436-z

Bozkurt, A., Xiao, J., Farrow, R., Bai, J. Y. H., Nerantzi, C., Moore, S.,...Asino, T. I. (2024). The Manifesto for Teaching and Learning in a Time of Generative AI: A Critical Collective Stance to Better Navigate the Future. *Open Praxis*, *16*(4), 487-513. https://doi.org/10.55982/openpraxis.16.4.777

Buijsman, S. (2024). Transparency for AI Systems: A Value-Based Approach. *Ethics and Information Technology*, *26*(2), 1-11. https://doi.org/10.1007/s10676-024-09770-w

Chan, C. K. Y. (2023). A Comprehensive AI Policy Education Framework for University Teaching and Learning. *International Journal of Educational Technology in Higher Education*, *20*(1), 1-25. https://doi.org/10.1186/s41239-023-00408-3

Corbin, T., Bearman, M., Boud, D., & Dawson, P. (2025). The Wicked Problem of AI and Assessment. *Assessment & Evaluation in Higher Education*, 1-17. https://doi.org/10.1080/02602938.2025.2553340

Dignum, V. (2019). *Responsible Artificial Intelligence: How to Develop and Use AI in a Responsible Way*. Springer Cham. https://doi.org/10.1007/978-3-030-30371-6

Dignum, V., Baldoni, M., Baroglio, C., Caon, M., Chatila, R., Dennis, L.,...Wildt, T. d. (2018). Ethics by Design: Necessity or Curse? *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*. https://doi.org/10.1145/3278721.3278745

Durán, J. M., & Pozzi, G. (2025). Trust and Trustworthiness in AI. *Philosophy & Technology*, *38*(1), 1-31. https://doi.org/10.1007/s13347-025-00843-2

Fanni, R., Steinkogler, V. E., Zampedri, G., & Pierson, J. (2023). Enhancing Human Agency Through Redress in Artificial Intelligence Systems. *AI & SOCIETY*, *38*(2), 537-547. https://doi.org/10.1007/s00146-022-01454-7

Felzmann, H., Fosch-Villaronga, E., Lutz, C., & Tamò-Larrieux, A. (2020). Towards Transparency by Design for Artificial Intelligence. *Science and Engineering Ethics*, *26*(6), 3333-3361. https://doi.org/10.1007/s11948-020-00276-4

Fu, Y., & Weng, Z. (2024). Navigating the Ethical Terrain of AI in Education: A Systematic Review on Framing Responsible Human-Centered AI Practices. *Computers and Education: Artificial Intelligence*, *7*, 1-20. https://doi.org/10.1016/j.caeai.2024.100306

Garcia, M. B. (2025). ChatGPT as an Academic Writing Tool: Factors Influencing Researchers' Intention to Write Manuscripts Using Generative Artificial Intelligence. *International Journal of Human-Computer Interaction*. https://doi.org/10.1080/10447318.2025.2499158

Garcia, M. B., Rosak-Szyrocka, J., Yılmaz, R., Metwally, A. H. S., Acut, D. P., Ofosu-Ampong, K.,...Bozkurt, A. (2025). Rethinking Educational Assessment in the Age of Generative AI: Actionable Strategies to Mitigate Academic Dishonesty. In *Pitfalls of AI Integration in Education: Skill Obsolescence, Misuse, and Bias* (pp. 1-24). IGI Global. https://doi.org/10.4018/979-8-3373-0122-8.ch001

Hasanah, N. A., Aziza, M. R., Junikhah, A., Arif, Y. M., & Garcia, M. B. (2025). Navigating the Use of AI in Engineering Education Through a Systematic Review of Technology, Regulations, and Challenges. In *Pitfalls of AI Integration in Education: Skill Obsolescence, Misuse, and Bias*. IGI Global. https://doi.org/10.4018/979-8-3373-0122-8.ch016

Henrique, B. M., & Santos, E. (2024). Trust in Artificial Intelligence: Literature Review and Main Path Analysis. *Computers in Human Behavior: Artificial Humans*, *2*(1), 1-13. https://doi.org/10.1016/j.chbah.2024.100043

Holmes, W., Porayska-Pomsta, K., Holstein, K., Sutherland, E., Baker, T., Shum, S. B.,...Koedinger, K. R. (2022). Ethics of AI in Education: Towards a Community-Wide Framework. *International Journal of Artificial Intelligence in Education*, *32*(3), 504-526. https://doi.org/10.1007/s40593-021-00239-1

Izquierdo-Álvarez, V., & Jimeno-Postigo, C. (2025). Challenges and Opportunities of Integrating Generative Artificial Intelligence in Higher Education: A Systematic Review. In *Pitfalls of AI Integration in Education: Skill Obsolescence, Misuse, and Bias*. IGI Global. https://doi.org/10.4018/979-8-3373-0122-8.ch017

Kamali, J., Alpat, M. F., & Bozkurt, A. (2024). AI Ethics as a Complex and Multifaceted Challenge: Decoding Educators' AI Ethics Alignment Through the Lens of Activity Theory. *International Journal of Educational Technology in Higher Education*, *21*(1), 1-20. https://doi.org/10.1186/s41239-024-00496-9

Kelly, S., Kaye, S.-A., & Oviedo-Trespalacios, O. (2023). What Factors Contribute to the Acceptance of Artificial Intelligence? A Systematic Review. *Telematics and Informatics*, *77*, 1-33. https://doi.org/10.1016/j.tele.2022.101925

Krakowski, S. (2025). Human-AI Agency in the Age of Generative AI. *Information and Organization*, *35*(1), 1-25. https://doi.org/10.1016/j.infoandorg.2025.100560

Larsson, S., & Heintz, F. (2020). Transparency in Artificial Intelligence. *Internet Policy Review*, *9*(2), 1-16. https://doi.org/10.14763/2020.2.1469

Legaspi, R., Xu, W., Konishi, T., Wada, S., Kobayashi, N., Naruse, Y., & Ishikawa, Y. (2024). The Sense of Agency in Human–AI Interactions. *Knowledge-Based Systems*, *286*, 1-16. https://doi.org/10.1016/j.knosys.2023.111298

Luo, J. (2024). A Critical Review of GenAI Policies in Higher Education Assessment: A Call to Reconsider the "Originality" of Students' Work. *Assessment & Evaluation in Higher Education*, *49*(5), 651-664. https://doi.org/10.1080/02602938.2024.2309963

Martínez-Requejo, S., Redondo-Duarte, S., Jiménez-García, E., & Ruiz-Lázaro, J. (2025). Technoethics and the Use of Artificial Intelligence in Educational Contexts: Reflections on Integrity, Transparency, and Equity. In *Pitfalls of AI Integration in Education: Skill Obsolescence, Misuse, and Bias*. IGI Global. https://doi.org/10.4018/979-8-3373-0122-8.ch010

Memarian, B., & Doleck, T. (2023). Fairness, Accountability, Transparency, and Ethics (FATE) in Artificial Intelligence (AI) and higher education: A Systematic Review. *Computers and Education: Artificial Intelligence*, *5*, 1-12. https://doi.org/10.1016/j.caeai.2023.100152

Nazaretsky, T., Mejia-Domenzain, P., Swamy, V., Frej, J., & Käser, T. (2025). The Critical Role of Trust in Adopting AI-Powered Educational Technology for Learning: An Instrument for Measuring Student Perceptions. *Computers and Education: Artificial Intelligence*, *8*, 1-16. https://doi.org/10.1016/j.caeai.2025.100368

Nguyen, A., Ngo, H. N., Hong, Y., Dang, B., & Nguyen, B.-P. T. (2023). Ethical Principles for Artificial Intelligence in Education. *Education and Information Technologies*, *28*(4), 4221-4241. https://doi.org/10.1007/s10639-022-11316-w

Ouyang, F., & Jiao, P. (2021). Artificial Intelligence in Education: The Three Paradigms. *Computers and Education: Artificial Intelligence*, *2*, 1-6. https://doi.org/10.1016/j.caeai.2021.100020

Radanliev, P. (2025). AI Ethics: Integrating Transparency, Fairness, and Privacy in AI Development. *Applied Artificial Intelligence*, *39*(1), 1-41. https://doi.org/10.1080/08839514.2025.2463722

Shata, A., & Hartley, K. (2025). Artificial Intelligence and Communication Technologies in Academia: Faculty Perceptions and the Adoption of Generative AI. *International Journal of Educational Technology in Higher Education*, *22*(1), 1-22. https://doi.org/10.1186/s41239-025-00511-7

Stanoyevitch, A. (2024). Online Assessment in the Age of Artificial Intelligence. *Discover Education*, *3*(1), 1-12. https://doi.org/10.1007/s44217-024-00212-9

Swiecki, Z., Khosravi, H., Chen, G., Martinez-Maldonado, R., Lodge, J. M., Milligan, S.,...Gašević, D. (2022). Assessment in the Age of Artificial Intelligence. *Computers and Education: Artificial Intelligence*, *3*, 1-10. https://doi.org/10.1016/j.caeai.2022.100075

Wachira, P. W., Liu, X., & Koc, S. (2025). *Educational Assessments in the Age of Generative AI*. IGI Global. https://doi.org/10.4018/979-8-3693-6351-5

Wanner, J., Herm, L.-V., Heinrich, K., & Janiesch, C. (2022). The Effect of Transparency and Trust on Intelligent System Acceptance: Evidence From a User-Based Study. *Electronic Markets*, *32*(4), 2079-2102. https://doi.org/10.1007/s12525-022-00593-5

Wiese, L. J., Patil, I., Schiff, D. S., & Magana, A. J. (2025). AI Ethics Education: A Systematic Literature Review. *Computers and Education: Artificial Intelligence*, *8*, 1-22. https://doi.org/https://doi.org/10.1016/j.caeai.2025.100405

Xiao, J., Bozkurt, A., Nichols, M., Pazurek, A., Stracke, C. M., Bai, J. Y. H.,...Themeli, C. (2025). Venturing into the Unknown: Critical Insights into Grey Areas and Pioneering Future Directions in Educational Generative AI Research. *TechTrends*, 1-16. https://doi.org/10.1007/s11528-025-01060-6

Yan, Y., & Liu, H. (2024). Ethical Framework for AI Education Based on Large Language Models. *Education and Information Technologies*, 1-19. https://doi.org/10.1007/s10639-024-13241-6

RELATED RESEARCH

Journal Article

Human–AI Interaction in a Socio-Educational Metaverse: Insights from a Developmental Evaluation of AI Avatars

Garcia, M. B. (2026). *Interactive Learning Environments*, 1-18. https://manuelgarcia.info/publication/ai-educational-metaverse

Book Chapter

Ethical Dilemmas and Practical Strategies for Leveraging Generative AI in STEM Assessment in Higher Education

Acut, D. P., Malayao, S. O., Buan, A. T., Caparoso, J. K. V., Mangubat, J. C., & Garcia, M. B. (2026). In Lahby, M., Schaeffer, S.E., Maleh, Y., & Paliktzoglou, V. (Eds.), *Generative AI in Higher Education Assessment* (pp. 229–262). Springer Nature. https://manuelgarcia.info/publication/generative-ai-in-stem-assessment

Conference Paper

Less Watching, Less Learning? Investigating the Immediate Effects of AI-Generated Summaries in Video-Based Learning

Garcia, M. B., Yousef, A. M. F., Happonen, A., & Crompton, H. (2026). In *2025 2nd International Conference on Artificial Intelligence and Teacher Education* (pp. 176-182). https://manuelgarcia.info/publication/ai-video-summaries

LET'S COLLABORATE!

If you are looking for research collaborators, please do not hesitate to contact me at mbgarcia@feutech.edu.ph.

ABOUT THE CORRESPONDING AUTHOR:

Manuel B. Garcia is a professor of information technology and the founding director of the Educational Innovation and Technology Hub (EdITH) at FEU Institute of Technology, Manila, Philippines. His interdisciplinary research interest includes topics that, individually or collectively, cover the disciplines of education and information technology. He is a licensed professional teacher and a proud member of the National Research Council of the Philippines – an attached agency to the country's Department of Science and Technology (DOST-NRCP).