Toward ethical and credible AI-assisted assessment: Reframing authenticity, transparency, and trust in higher education

Clark Dominic Alipasa, (https://orcid.org/0000-0003-3459-6212), De La Salle University-Dasmarinas, Philippines; Mariano Thomas Ramirez, (https://orcid.org/0000-0002-8497-1204), National University-Dasmarinas, Philippines

Abstract

Artificial intelligence (AI), particularly generative systems, is reshaping higher education assessment by challenging traditional understandings of authenticity, authorship, and credibility. While AI offers affordances such as personalized feedback, instructional efficiency, and enhanced learning support, it also introduces significant risks, including hallucinated outputs, misinformation, and reduced transparency in knowledge production. These developments complicate the evaluation of student performance because AI-generated content may appear academically coherent and credible while lacking epistemic validity. Drawing on scholarship in human–AI interaction, credibility theory, and algorithmic transparency, this paper argues that assessment must be reconceptualized as an interpretive and interactional process shaped by both human and machine agency. In response, it proposes a conceptual framework integrating authenticity, trust, and transparency.

Keywords: Artificial intelligence, higher education assessment, authenticity, credibility, transparency, trust

1. Introduction

Artificial intelligence (AI), particularly generative systems such as large language models, is rapidly reshaping the epistemic and pedagogical foundations of higher education assessment. These technologies are no longer peripheral instructional tools but are increasingly embedded in how students generate, revise, and submit academic work. Their capacity to produce coherent, contextually appropriate, and human-like outputs has fundamentally altered the conditions under which learning evidence is created and evaluated. As a result, traditional assumptions regarding authorship, originality, and individual cognitive effort are becoming increasingly unstable, requiring a reconsideration of what constitutes valid academic performance in AI-mediated environments (Dwivedi et al., 2023; Fui-Hoon Nah et al., 2023).

This transformation is particularly significant for authentic assessment, which has long functioned as a corrective to standardized and decontextualized evaluation practices. Authentic assessment emphasizes complex, real-world tasks that require students to demonstrate applied understanding, critical reasoning, and contextualized problem-solving. However, generative AI complicates this pedagogical foundation by introducing systems capable of simulating many of the cognitive processes that assessments are designed to evaluate. When AI systems can generate essays, analyses, and solutions that closely resemble competent student work, the distinction between demonstrated understanding and algorithmically generated output becomes increasingly difficult to sustain.

Beyond concerns of authorship, the integration of AI into assessment introduces a deeper epistemic challenge involving credibility and truth validation. Generative models are known to produce "hallucinations," or outputs that are linguistically fluent yet factually inaccurate or fabricated. Such outputs may appear authoritative, particularly when embedded within polished academic discourse, making inaccuracies difficult for students and educators to detect (Alkaissi & McFarlane, 2023). Simultaneously, the rapid circulation of AI-generated content contributes to broader concerns regarding misinformation and epistemic distortion within digital knowledge systems (Monteith et al., 2024). Consequently, assessment is no longer concerned solely with what students know but also with whether presented knowledge can be epistemically trusted.

These concerns are intensified by the structural opacity of AI systems. Most generative models operate through highly complex and non-transparent architectures that provide limited explainability regarding how outputs are produced. This "black box" condition constrains users' ability to meaningfully interrogate the origins, reliability, or validity of generated content. Although transparency is frequently proposed as a mechanism for algorithmic accountability, research suggests that visibility alone does not necessarily produce interpretability or meaningful control, particularly for non-technical users (Ananny & Crawford, 2018; Larsson & Heintz, 2020). Within educational contexts, this creates conditions in which assessment judgments are made under persistent epistemic uncertainty.

From a theoretical perspective, these developments can be situated within frameworks of human–AI interaction and credibility assessment. The concept of machine agency suggests that users increasingly attribute communicative and cognitive capacities to AI systems, treating them as active participants in meaning-making processes rather than passive technological tools (Sundar, 2020). At the same time, credibility research demonstrates that information evaluation is often shaped by heuristic cues such as fluency, coherence, and perceived authority rather than systematic verification processes (Metzger, 2007; Sundar, 2007). Generative AI intensifies these tendencies by producing outputs that strongly activate credibility heuristics, thereby increasing the risk of overreliance and misjudgment of epistemic quality.

Taken together, these developments reveal a fundamental tension within contemporary assessment systems: the need to preserve authenticity, credibility, and trust while simultaneously engaging with technologies capable of replicating and obscuring those same qualities. This tension cannot be resolved through minor procedural adjustments because it reflects a deeper transformation in the epistemology of educational evaluation itself. Instead, it necessitates a reconceptualization of assessment as an interpretive and interactive process shaped by human cognition, algorithmic systems, and institutional norms.

In response, this paper argues that authentic assessment in the age of AI must be reframed as a negotiated and interpretive process rather than a purely human-centered demonstration of knowledge. Drawing on interdisciplinary scholarship from education, communication, and information science, the study examines how generative AI reshapes conditions of credibility and trust in assessment. Building on this synthesis, the paper proposes a conceptual framework that integrates authenticity, transparency, and trust as interdependent dimensions of AI-assisted assessment. This framework seeks to contribute to ongoing debates regarding how higher

education can preserve epistemic integrity while adapting to the realities of AI-mediated knowledge production.

2. Authentic Assessment in the Age of Artificial Intelligence

Authentic assessment has long been regarded as a response to the limitations of traditional standardized evaluation methods in higher education. Rather than emphasizing decontextualized recall of information, authentic assessment focuses on tasks that reflect real-world application, requiring students to demonstrate higher-order thinking, problem-solving, and contextualized understanding. This paradigm assumes that learning is best evaluated through meaningful performance in which students actively construct and apply knowledge in complex situations. However, the emergence of artificial intelligence (AI), particularly generative systems, disrupts this foundational assumption by introducing technologies capable of producing outputs that closely resemble human-authored work. Consequently, the integration of AI into educational contexts necessitates a reconsideration of authenticity as a central principle of assessment (Donghee Shin, 2022; Donghee Shin & Park, 2019).

Beyond its pedagogical purpose, authentic assessment is closely tied to issues of credibility and trust. The validity of assessment outcomes depends on the assumption that student outputs genuinely reflect their knowledge, abilities, and intellectual engagement. In AI-mediated environments, however, this assumption becomes increasingly unstable. Generative AI systems can assist, augment, or fully produce academic outputs, making it difficult to determine the extent of actual student contribution. This shift challenges not only assessment design but also the epistemological foundations of evaluation, as educators must increasingly account for hybrid forms of authorship involving both human and machine participation (S. Shyam Sundar, 2020; Michael J. Metzger, 2007).

2.1 Authenticity and Its Pedagogical Foundations

The concept of authenticity in assessment is grounded in constructivist perspectives of learning, where knowledge is actively constructed through engagement with meaningful tasks and environments. Authentic assessment practices are intended to mirror professional or real-world contexts, enabling students to demonstrate not only content mastery but also transferable skills and applied understanding. Such practices prioritize complexity, contextualization, and learner agency, positioning students as active participants in the assessment process rather than passive recipients of evaluation.

Even prior to the rise of AI, authentic assessment faced both conceptual and practical challenges. Concerns regarding subjectivity, scalability, consistency, and evaluative reliability have long complicated its implementation. These challenges become even more pronounced within digital environments, where the boundaries between original and mediated work are increasingly difficult to distinguish. The introduction of AI intensifies these concerns by enabling the rapid generation of contextually appropriate yet potentially superficial, misleading, or machine-mediated outputs, thereby further complicating judgments of authenticity and evaluative validity (Donghee Shin, 2023; Hee Eun Park, 2024).

2.2 AI as an Affordance and Constraint in Assessment

From an affordances perspective, AI technologies possess significant potential to enhance assessment practices in higher education. Generative AI systems can support personalized feedback, scaffold learning processes, and assist students in idea development, thereby contributing to more adaptive and learner-centered educational experiences. These capabilities align with broader trends in educational technology emphasizing flexibility, responsiveness, and individualized learning design (Yogesh K. Dwivedi et al., 2023; Fui-Hoon Nah et al., 2023).

At the same time, AI introduces constraints that challenge the integrity and validity of assessment. The same technologies that enable support and instructional augmentation may also encourage over-reliance, reduce cognitive engagement, or generate outputs that obscure a learner's actual level of understanding. This duality reflects the "double-edged sword" nature of generative AI, in which pedagogical affordances are inseparable from accompanying epistemic and evaluative risks (Hee Eun Park, 2024; Donghee Shin, 2022). Consequently, assessment practices must navigate the tension between leveraging the benefits of AI integration and preserving meaningful indicators of authentic learning and intellectual performance.

2.3 Credibility, Trust, and the Problem of AI-Generated Knowledge

A central concern in AI-assisted assessment involves the credibility of generated outputs. Credibility, understood as the perceived accuracy, reliability, and trustworthiness of information, plays a significant role in how students and educators evaluate academic work. Within digital environments, credibility judgments are frequently shaped by heuristic cues such as presentation quality, fluency, and perceived authority rather than systematic verification of content accuracy (Michael J. Metzger et al., 2010; S. Shyam Sundar, 2007).

Generative AI complicates these evaluative processes by producing outputs that are linguistically coherent and contextually appropriate yet not necessarily factually accurate. The phenomenon of AI "hallucinations," in which systems generate false, fabricated, or misleading information, presents substantial risks for assessment, particularly when such inaccuracies are difficult to detect (Hussam Alkaissi & McFarlane, 2023; Yogesh K. Dwivedi et al., 2023). As a result, the appearance of credibility may diverge from actual epistemic validity, thereby undermining trust in assessment outcomes and educational evaluation more broadly.

2.4 Transparency, Accountability, and Human–AI Interaction

The challenges surrounding credibility are closely connected to broader issues of transparency and accountability within AI systems. Many generative models operate with limited visibility into their internal processes, making it difficult for users to understand how outputs are produced or meaningfully evaluate their reliability. Although transparency is frequently proposed as a mechanism for improving algorithmic accountability, scholars argue that transparency alone is insufficient because users may lack the expertise or motivation necessary to interpret complex algorithmic systems effectively (Michael Ananny & Kate Crawford, 2018; Stefan Larsson & Heintz, 2020).

From the perspective of human–AI interaction, users engage with AI systems through ongoing processes of interpretation, negotiation, and sensemaking. These interactions shape how AI-generated outputs are understood, trusted, and incorporated into educational practice. The concept of machine agency highlights the growing tendency of users to attribute autonomy, intentionality, and communicative capacity to AI systems, which subsequently influences how generated outputs are evaluated and legitimized (S. Shyam Sundar, 2020; Donghee Shin et al., 2024). Within assessment environments, this suggests that authenticity and credibility are not fixed or purely objective properties but are co-constructed through dynamic interactions between human users and AI systems.

3. Artificial Intelligence in Contemporary Assessment Practices

The integration of artificial intelligence (AI) into higher education assessment is no longer speculative but increasingly embedded within everyday academic practice. From automated feedback systems to generative text production, AI is reshaping how assessment is designed, completed, and evaluated. These developments extend beyond efficiency gains, fundamentally altering both the nature of student engagement and the evidentiary foundations upon which learning is assessed. As AI systems become more accessible and sophisticated, their role is shifting from peripheral support technologies to active participants in knowledge construction, thereby redefining the boundaries of assessment itself (Yogesh K. Dwivedi et al., 2023; Fui-Hoon Nah et al., 2023).

At the same time, the adoption of AI within assessment remains uneven and context-dependent, shaped by institutional policies, disciplinary cultures, and user perceptions. While some educators integrate AI as a pedagogical resource, others perceive it as a threat to academic integrity and authentic learning. This divergence reflects broader uncertainties regarding how AI should be positioned within educational systems, particularly in relation to authorship, originality, and the evaluation of learning outcomes. Consequently, understanding AI in assessment requires not only a technical perspective but also a socio-cognitive framework that considers how users interpret, negotiate, and assign meaning to AI's role in academic work (Donghee Shin, 2022; Bettina K. Waruwu et al., 2021).

3.1 AI as a Tool for Personalization and Feedback

One of the most frequently cited affordances of AI in education is its capacity to provide personalized feedback at scale. AI-driven systems can analyze student responses, identify patterns, and generate tailored feedback that supports learning progression and instructional adaptation. This capability addresses long-standing challenges in higher education, particularly in large classes where individualized feedback is constrained by time and institutional resources. By automating aspects of evaluation and instructional response, AI enables more timely and adaptive interactions between learners and educational content (Fui-Hoon Nah et al., 2023; Yogesh K. Dwivedi et al., 2023).

However, the effectiveness of AI-generated feedback depends significantly on how it is perceived, interpreted, and utilized by students. Research indicates that users' trust in AI systems strongly influences their willingness to engage with and act upon automated feedback. Factors

such as perceived credibility, transparency, and human-likeness play important roles in shaping these perceptions and interactions (S. Shyam Sundar, 2020; Bin Liu, 2021). Consequently, although AI offers substantial opportunities for enhancing feedback mechanisms, its educational impact remains mediated by complex psychological and interactional dynamics.

3.2 Generative AI and the Transformation of Student Work

Generative AI represents a more substantial transformation in assessment practices by enabling the production of complete academic outputs, including essays, reports, analyses, and problem solutions. Unlike earlier educational technologies that primarily supported isolated components of learning, generative systems can simulate higher-order cognitive processes, raising important questions regarding the extent to which submitted work genuinely reflects student understanding. This capability disrupts traditional assumptions about authorship and complicates the interpretation of assessment outcomes.

The widespread accessibility of generative AI tools has also altered student learning behaviors and academic strategies. Emerging research suggests that students increasingly use AI not only for assistance but also for brainstorming, drafting, revising, and organizing academic work, effectively integrating these systems into everyday learning workflows (Lina I. D. Faruk et al., 2023; Yogesh K. Dwivedi et al., 2023). While such practices may improve efficiency and productivity, they also blur the distinction between meaningful learning and technological outsourcing, making it increasingly difficult to separate genuine understanding from AI-supported performance. This transformation necessitates a reconsideration of what constitutes valid evidence of learning within AI-mediated educational environments.

3.3 Human–AI Collaboration and Hybrid Authorship

Rather than conceptualizing AI solely as either a threat or a tool, recent scholarship increasingly emphasizes the collaborative dimensions of human–AI interaction. Within this perspective, AI is understood as a partner that augments human capabilities and enables new forms of creativity, reasoning, and problem-solving. This shift aligns with broader discussions of AI as a co-creator, where outputs emerge through iterative interactions between human input and machine-generated responses (Donghee Shin, 2023; S. Shyam Sundar, 2020).

However, the concept of collaboration introduces significant complexities for assessment, particularly regarding agency, authorship, and responsibility. If academic outputs are co-produced through human–AI interaction, traditional evaluative criteria such as originality, individual effort, and independent cognition may no longer be sufficient. Instead, assessment frameworks must increasingly account for processes of interaction, decision-making, and critical engagement with AI-generated content. This shift requires movement away from purely product-oriented evaluation toward process-oriented assessment, where greater emphasis is placed on how students use AI systems rather than simply whether they use them.

3.4 Emerging Tensions in AI-Integrated Assessment

The integration of AI into assessment generates several interconnected tensions that challenge existing educational paradigms. First, there is a tension between efficiency and depth, as AI enables rapid production of outputs while potentially reducing opportunities for sustained cognitive engagement and reflective thinking. Second, there is a tension between accessibility and integrity, as AI technologies democratize access to knowledge and academic support while simultaneously increasing risks of misuse, dependency, and over-reliance. Third, there is a tension between innovation and regulation, as educational institutions struggle to balance technological adoption with the preservation of academic standards and evaluative legitimacy.

These tensions are not merely technical but reflect broader epistemological and pedagogical concerns regarding the future of educational evaluation. They underscore the need for assessment frameworks capable of accommodating the complexities of AI-mediated learning while preserving core educational values such as authenticity, credibility, and intellectual rigor. Importantly, these tensions also establish the foundation for deeper concerns related to misinformation, trust, and epistemic reliability, which will be examined in subsequent sections (Sean T. Monteith et al., 2024; Donghee Shin et al., 2024).

4. Risks, Credibility, and Misinformation in AI-Assisted Assessment

The integration of generative artificial intelligence into educational assessment introduces a significant epistemic disruption in how credibility is constructed, evaluated, and sustained. Unlike earlier digital technologies that primarily supported information retrieval, organization, or editing, generative AI systems actively produce fluent, contextually coherent, yet potentially inaccurate content. This development creates conditions in which surface-level linguistic quality no longer reliably corresponds to epistemic validity or factual accuracy. A major concern involves AI "hallucinations," where systems generate fabricated, misleading, or false information that nevertheless appears authoritative and convincing to users (Alkaissi & McFarlane, 2023; Dwivedi et al., 2023). Within assessment environments, this creates a serious evaluative problem: student outputs may appear academically sophisticated while remaining epistemically unstable, thereby weakening the reliability and integrity of assessment systems.

4.1 AI Hallucinations and the Distortion of Epistemic Validity

One of the most significant risks in AI-assisted assessment is the phenomenon of hallucinated content, in which generative systems produce information lacking grounding in verifiable evidence or factual data. These outputs are not random mistakes but structurally plausible statements generated through probabilistic language-modeling processes. Consequently, hallucinated content frequently evades detection because of its coherence, fluency, and academically polished presentation. Alkaissi and McFarlane (2023) emphasize that such hallucinations have direct implications for scientific and academic writing, particularly when users interpret linguistic fluency as evidence of correctness. Similarly, Dwivedi et al. (2023) argue that generative AI introduces new epistemic vulnerabilities into knowledge production systems by separating textual sophistication from factual accuracy.

Within assessment contexts, this development creates a serious validity problem. Educators are no longer evaluating only student understanding but also the epistemic integrity of machine-

assisted outputs. The challenge lies in the fact that hallucinated responses are often difficult to distinguish from accurate information without extensive external verification. This destabilizes traditional assessment assumptions in which clarity, coherence, and structural organization are treated as indicators of understanding. Consequently, AI hallucinations introduce a concealed layer of epistemic distortion that undermines the reliability of grading, feedback, and educational evaluation.

4.2 Credibility Formation in AI-Mediated Environments

Credibility in AI-assisted assessment cannot be understood as a fixed property of information but rather as a perception shaped by cognitive, contextual, and interactional processes. Research on online credibility demonstrates that individuals frequently rely on heuristic cues such as fluency, structure, and perceived authority when evaluating information quality (Metzger, 2007; Metzger et al., 2010). Generative AI systems are especially effective at activating these heuristics because they consistently produce polished and contextually appropriate outputs that may incorrectly signal factual reliability.

At the same time, credibility judgments are influenced by expectancy-based mechanisms through which users interpret information according to prior beliefs, assumptions, and system expectations (Burgoon & Hale, 1988; Appelman & Sundar, 2016). In AI-mediated environments, such expectations are often amplified by the human-like communicative style of generative systems. As a result, students and educators may over-trust AI-generated outputs despite the absence of systematic verification. This creates a systemic risk in which perceived credibility diverges from actual epistemic reliability, particularly within high-stakes assessment contexts.

4.2.1 Heuristic Processing and the Illusion of Accuracy

A deeper dimension of the credibility problem lies in how users cognitively process AI-generated content. According to heuristic-systematic processing models, individuals frequently rely on cognitive shortcuts when confronted with complex, ambiguous, or time-sensitive information (Metzger et al., 2010; Koh & Sundar, 2010). Generative AI intensifies this tendency by producing outputs that are not only fluent but also stylistically aligned with academic conventions and discourse patterns.

This fluency creates what may be described as an “illusion of accuracy,” in which well-structured and polished text is mistakenly interpreted as factually valid. Appelman and Sundar (2016) demonstrate that message credibility is strongly influenced by presentation quality, while Koh and Sundar (2010) show that users commonly depend on superficial cues when evaluating online information. Within AI-assisted assessment, this suggests that students may accept AI-generated content without adequate critical interrogation, while educators may struggle to identify inaccuracies embedded within apparently sophisticated responses.

The implications are substantial. Credibility in AI-mediated assessment becomes less dependent on truth verification and increasingly shaped by perceived plausibility and rhetorical fluency. This shift destabilizes conventional academic evaluation criteria, which traditionally assume an alignment between coherence and correctness. Instead, AI introduces a growing disconnect

between form and epistemic substance, requiring new evaluative approaches capable of accounting for machine-generated persuasion effects.

4.3 Algorithmic Opacity and the Limits of Transparency

Beyond hallucinations and heuristic bias, credibility challenges are further intensified by the inherent opacity of generative AI systems. Most large language models function as highly complex "black boxes," in which relationships among input data, training processes, and generated outputs remain difficult to interpret. Ananny and Crawford (2018) argue that transparency, although frequently presented as a solution to algorithmic accountability, is fundamentally limited in its ability to generate meaningful understanding. Simply exposing system processes does not guarantee that users will be able to interpret, evaluate, or critically assess them effectively.

Similarly, Larsson and Heintz (2020) contend that algorithmic transparency must be understood relative to user knowledge, interpretive capacity, and contextual conditions. Within educational settings, both students and educators may lack the technical expertise necessary to meaningfully evaluate AI system behavior. This creates a significant paradox: even when AI systems are intentionally designed to be transparent, their outputs often remain epistemically opaque to most users.

As a consequence, transparency risks becoming performative rather than genuinely functional. It may provide the appearance of accountability without ensuring actual interpretability, comprehension, or control. Within assessment environments, this means that judgments regarding the credibility of AI-assisted outputs are frequently made without full visibility into how those outputs were generated, thereby further weakening trust in evaluation systems.

4.4 A Multi-Layered Credibility Crisis

Taken together, AI hallucinations, heuristic-based credibility judgments, and algorithmic opacity generate a multi-layered credibility crisis within assessment systems. At the content level, AI systems may produce inaccurate yet highly plausible information (Alkaissi & McFarlane, 2023). At the cognitive level, users may over-rely on fluency, structure, and presentation quality as indicators of truth and validity (Metzger et al., 2010; Appelman & Sundar, 2016). At the structural level, limitations in transparency restrict meaningful verification and interpretation of AI processes (Ananny & Crawford, 2018; Larsson & Heintz, 2020).

This convergence suggests that credibility in AI-assisted assessment cannot be conceptualized as a stable or singular construct. Instead, it must be understood as a dynamic outcome emerging through interactions among human cognition, algorithmic design, and institutional evaluation practices. The resulting instability of credibility establishes the foundation for the next critical issue: how transparency and trust may be reconstructed within AI-mediated educational environments.

5. Transparency, Trust, and Human–AI Interaction in Assessment

The credibility crisis introduced by generative artificial intelligence in assessment does not exist in isolation but is closely intertwined with broader issues of transparency, trust, and human–AI interaction. While Section 4 demonstrated how hallucinations, heuristic processing, and algorithmic opacity destabilize credibility, this section examines how these disruptions are managed, negotiated, or further intensified through user interaction with AI systems. Within contemporary educational environments, trust is no longer directed exclusively toward human evaluators or institutional structures but is increasingly distributed across both human and non-human agents. This shift requires a reconceptualization of trust as a relational and mediated construct shaped by algorithmic systems, user cognition, and contextual expectations (Sundar, 2020; Shin, 2023).

5.1 Algorithmic Transparency and Its Epistemic Limitations

Transparency has frequently been proposed as a foundational principle for ensuring accountability within algorithmic systems, including AI-assisted educational technologies. The underlying assumption is that making system processes visible enables users to better evaluate outputs, identify limitations, and make informed judgments. However, recent scholarship has critically challenged this assumption. Ananny and Crawford (2018) argue that transparency is inherently limited because it often produces visibility without comprehension. In other words, making systems observable does not necessarily render them understandable, interpretable, or meaningfully evaluable.

Larsson and Heintz (2020) further contend that algorithmic transparency must be understood relative to user capacity, institutional structures, and interpretive competencies. Within higher education settings, these limitations are particularly significant because students and educators may lack the technical literacy necessary to interpret model behavior and assess algorithmic reliability. Consequently, transparency becomes asymmetrical: AI systems may become increasingly visible in principle while remaining opaque in practice. This paradox constrains the extent to which transparency can function as an effective mechanism for restoring trust in educational assessment processes.

5.2 Trust Formation in Human–AI Interaction

Trust in AI-assisted assessment should not be understood as a fixed belief but as an evolving psychological and relational process shaped through interaction with intelligent systems. According to Sundar's (2020) framework of machine agency, users attribute varying degrees of autonomy, intentionality, and competence to AI systems depending on interaction experiences and system design cues. These attributions significantly influence how AI-generated outputs are perceived, evaluated, and integrated into decision-making processes.

Shin (2023) extends this perspective by arguing that algorithmic interactions are inherently interpretive, requiring users to engage continuously in processes of sensemaking in order to evaluate system reliability. Within assessment contexts, trust is therefore not pre-established but continually constructed and reconstructed through repeated interaction. Students may initially

trust AI-generated outputs because of their fluency, coherence, and apparent sophistication, yet this trust may diminish when inconsistencies or inaccuracies become apparent. Conversely, repeated positive experiences may strengthen reliance on AI systems even in the absence of rigorous verification. Trust thus emerges as a dynamic consequence of ongoing human–AI interaction rather than a stable evaluative judgment.

5.3 Machine Agency and the Redistribution of Epistemic Responsibility

One of the most significant conceptual transformations introduced by AI in assessment involves the increasing perception of machine agency. Rather than functioning merely as passive tools, AI systems are increasingly perceived as active contributors to knowledge production and academic work. Sundar (2020) characterizes this phenomenon as the "rise of machine agency," in which users attribute communicative and cognitive capacities to algorithmic systems based on their outputs and interactional behavior.

Endacott (2024) further demonstrates that users frequently engage in relational interpretations of AI systems, treating them as quasi-social actors within decision-making and meaning-making processes. Within educational assessment, this perception complicates traditional understandings of authorship, accountability, and intellectual responsibility. If AI systems contribute substantially to the generation of academic outputs, responsibility for accuracy, originality, and epistemic integrity becomes distributed across both human and machine actors.

This redistribution challenges foundational assumptions in higher education assessment, which have traditionally located epistemic responsibility exclusively with the student. In contrast, AI-mediated environments require a reconsideration of accountability structures in which responsibility becomes shared, negotiated, and at times fundamentally ambiguous.

5.4 From Transparency to Interpretive Accountability

Given the limitations of transparency and the complexity of trust formation, recent scholarship increasingly advocates a shift toward interpretive accountability rather than reliance on purely technical transparency. This perspective emphasizes how users interpret, negotiate, and make sense of AI-generated outputs within particular contexts instead of assuming that system visibility alone guarantees ethical or trustworthy use.

Shin et al. (2024) argue that algorithmic sensemaking is central to how users evaluate fairness, reliability, and legitimacy in AI systems. Similarly, Liu (2021) demonstrates that perceptions of uncertainty reduction in human–AI interaction are shaped not only by system transparency but also by perceived agency locus, relational cues, and interactional dynamics. Within assessment environments, this suggests that trust is co-produced through interpretation, interaction, and institutional framing rather than embedded exclusively within system architecture or technical design.

From this perspective, transparency should not be understood as an endpoint but as a starting condition for ongoing interpretive engagement. Educators and students must critically engage with AI outputs not merely as informational products but as mediated constructions requiring

active evaluation, contextualization, and interrogation. Such a shift is essential for developing more resilient assessment practices capable of accommodating the complexities introduced by AI integration.

5.5 Toward Relational Trust in AI-Assisted Assessment

The analysis presented in this section suggests that neither transparency nor traditional trust mechanisms are sufficient to resolve the challenges introduced by AI in assessment. Instead, trust must be understood as relational, dynamic, and co-constructed through ongoing human–AI interaction. Algorithmic systems do not simply provide information; they actively shape how information is interpreted, evaluated, and legitimized within educational environments.

When considered alongside the credibility disruptions discussed in Section 4, these dynamics reveal a broader epistemic transformation in higher education assessment. Trust is no longer anchored solely in human authority, institutional oversight, or traditional evaluative structures but is increasingly distributed across interconnected networks of human cognition and algorithmic mediation. Consequently, assessment design must move toward frameworks that explicitly recognize interactional processes, interpretive labor, and shared epistemic responsibility as central dimensions of AI-assisted educational evaluation.

6. Toward a Conceptual Framework for Ethical and Credible AI-Assisted Assessment

The preceding sections have demonstrated that artificial intelligence in higher education assessment introduces not a single disruption but a layered epistemic transformation involving authenticity, credibility, transparency, and trust. Rather than conceptualizing these constructs as isolated concerns, this paper argues that they operate as interdependent dimensions within a broader evaluative system shaped by ongoing human–AI interaction. In response, this section synthesizes these dimensions into a conceptual framework that repositions authentic assessment as a negotiated, interpretive, and technologically mediated process.

6.1 Reframing Authenticity, Transparency, and Trust as Interdependent Constructs

Traditional models of authentic assessment assume that validity is grounded in the alignment between student performance and real-world task representation. Within AI-mediated environments, however, this assumption becomes increasingly insufficient because assessment performance itself may be partially or fully co-constructed through interaction with generative systems. As demonstrated in earlier sections, AI-generated content introduces significant uncertainty regarding credibility and epistemic validity (Section 4) while simultaneously reshaping processes of trust formation, interpretation, and meaning-making (Section 5).

Accordingly, authenticity can no longer be defined solely in terms of task realism or student independence. Instead, authenticity must increasingly be understood as interactional authenticity, where meaning emerges through the ways students engage with, evaluate, negotiate, and integrate AI-generated outputs into academic work. Within this framework, transparency and trust do not function merely as external safeguards or regulatory mechanisms but as mediating

dimensions shaping how authenticity is interpreted, negotiated, and validated in practice (Ananny & Crawford, 2018; Sundar, 2020; Shin, 2023).

This reconceptualization positions assessment not as a fixed measurement of individual performance but as an evolving ecosystem of negotiated meaning shaped through interactions among human cognition, algorithmic systems, and institutional expectations.

6.2 The Conceptual Framework: The Triadic Assessment Integrity Model

Building upon this interdisciplinary synthesis, the paper proposes the Triadic Assessment Integrity Model (TAIM), which integrates three interdependent dimensions:

- Authenticity (A) – the extent to which assessment reflects meaningful engagement with knowledge, including AI-mediated processes
- Transparency (T) – the degree to which AI system behavior and output generation can be understood, interpreted, or critically interrogated
- Trust (T*) – the relational and cognitive judgment of credibility formed through human–AI interaction and institutional framing

These dimensions interact dynamically rather than hierarchically. Authenticity is influenced by trust in AI-generated outputs, trust is shaped by perceived transparency, and transparency affects how authenticity is interpreted and evaluated. Importantly, breakdown within any one dimension destabilizes the integrity of the broader assessment system.

The TAIM framework extends prior scholarship on machine agency and algorithmic sensemaking (Sundar, 2020; Shin et al., 2024) by conceptualizing assessment not simply as an output-evaluation process but as a continuous interpretive cycle emerging through interaction between human users and AI systems.

6.3 Implications for Assessment Design and Pedagogy

The TAIM framework carries several implications for the redesign of assessment practices in higher education.

First, assessment tasks must move away from exclusively product-centered evaluation toward process-visible forms of assessment in which students document how AI systems were used, questioned, interpreted, and integrated into their academic work. Such an approach aligns with the need to evaluate interpretive engagement and epistemic reasoning rather than merely the production of polished textual outputs.

Second, educators must incorporate credibility literacy into assessment criteria and instructional practices. Given the risks associated with hallucinations, misinformation, and epistemic distortion (Alkaissi & McFarlane, 2023; Monteith et al., 2024), students should be evaluated not only on correctness but also on their ability to critically verify, contextualize, and interrogate AI-generated information.

Third, institutions must move beyond simplistic transparency mandates and instead cultivate interpretive transparency practices in which students and educators are trained to critically engage with AI outputs rather than passively relying on system disclosures or surface-level explanations (Ananny & Crawford, 2018; Larsson & Heintz, 2020).

Finally, trust itself must be treated as an educational outcome. Students should not be encouraged either to uncritically trust or categorically reject AI systems. Rather, they should develop calibrated trust grounded in contextual judgment, verification practices, and critical evaluation of AI-generated content.

6.4 Reconstructing Assessment in the Age of AI

This paper has argued that the integration of artificial intelligence into higher education assessment necessitates a fundamental reconceptualization of authenticity, credibility, transparency, and trust. Rather than treating AI as either a peripheral instructional tool or solely as a threat to academic integrity, it must be understood as an active participant in the epistemic construction of assessment outcomes.

By synthesizing interdisciplinary perspectives from human–AI interaction, credibility theory, and algorithmic transparency, this study developed the Triadic Assessment Integrity Model as a conceptual framework for understanding AI-assisted assessment. The model emphasizes that assessment integrity can no longer be guaranteed solely through assumptions of individual authorship or independent cognitive production. Instead, integrity emerges through interactions among human cognition, machine agency, interpretive engagement, and institutional structures.

Ultimately, the future of assessment in higher education will depend not on resisting artificial intelligence but on designing evaluative systems capable of critically and ethically engaging with it. This transformation requires a shift away from policing authenticity toward cultivating interpretive, credibility-aware, and ethically grounded assessment practices that reflect the realities of an increasingly AI-mediated knowledge ecosystem.

7. Conclusion

The rapid integration of generative artificial intelligence into higher education assessment has introduced profound epistemic, pedagogical, and evaluative challenges extending far beyond concerns about academic dishonesty or technological disruption. As this paper has demonstrated, AI systems fundamentally reshape how authenticity, credibility, transparency, and trust are constructed and interpreted within assessment environments. Generative AI not only alters the production of academic work but also destabilizes traditional assumptions regarding authorship, originality, and epistemic validity (Dwivedi et al., 2023; Fui-Hoon Nah et al., 2023). At the same time, hallucinated outputs, algorithmic opacity, and heuristic-based credibility judgments complicate the reliability of assessment systems and challenge conventional approaches to educational evaluation (Alkaissi & McFarlane, 2023; Metzger et al., 2010).

Drawing on interdisciplinary scholarship from human–AI interaction, credibility theory, and algorithmic transparency, this study argued that assessment can no longer be understood as a purely human-centered process of knowledge demonstration. Instead, assessment must

increasingly be conceptualized as a negotiated and interpretive interaction among students, educators, algorithmic systems, and institutional structures (Sundar, 2020; Shin, 2023). In response to these transformations, the paper proposed the Triadic Assessment Integrity Model (TAIM), which positions authenticity, transparency, and trust as interdependent dimensions shaping the integrity of AI-assisted assessment practices. This framework highlights that assessment integrity is not guaranteed solely through individual authorship or technological regulation but emerges through ongoing processes of interpretation, verification, and relational trust formation.

The implications of this reconceptualization are significant for higher education policy and pedagogy. Institutions must move beyond simplistic approaches that either prohibit or uncritically adopt AI technologies. Instead, educators and students must develop interpretive, credibility-aware, and ethically grounded assessment practices capable of critically engaging with AI-generated content (Ananny & Crawford, 2018; Larsson & Heintz, 2020). This includes emphasizing process-oriented assessment, credibility literacy, and calibrated trust rather than relying exclusively on traditional measures of originality or textual production. As AI systems continue to evolve, higher education will increasingly require evaluative frameworks recognizing the complex interplay among human cognition, machine agency, and institutional accountability.

Ultimately, the future of assessment in higher education will depend not on resisting artificial intelligence but on reimagining assessment systems that preserve epistemic integrity while adapting to the realities of AI-mediated knowledge production. Such a shift requires movement away from static notions of authenticity toward more dynamic, relational, and interpretive models of educational evaluation capable of addressing the complexities of emerging human–AI learning environments (Shin et al., 2024; Sundar, 2020).

References

Alkaissi, H., & McFarlane, S. I. (2023). Artificial hallucinations in ChatGPT: Implications in scientific writing. *Cureus, 15*(2), e35179.

Ananny, M., & Crawford, K. (2018). Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. *New Media & Society, 20*(3), 973–989.

Appelman, A., & Sundar, S. S. (2016). Measuring message credibility: Construction and validation of an exclusive scale. *Journalism & Mass Communication Quarterly, 93*(1), 59–79.

Bucher, T. (2017). 'Machines don't have instincts': Articulating the computational in journalism. *New Media & Society, 19*(6), 918–933.

Burgoon, J. K. (2015). Expectancy violations theory. In C. R. Berger et al. (Eds.), *The international encyclopedia of interpersonal communication* (pp. 1–9). Wiley.

Burgoon, J. K., & Hale, J. (1988). Nonverbal expectancy violations: Model elaboration and application to immediacy behaviors. *Communication Monographs, 55*(1), 58–79.

Chan, M. (2022). News literacy, fake news recognition, and authentication behaviors after exposure to fake news on social media. *New Media & Society, 26*, 4669–4688.

Chen, M., Liu, F., & Lee, Y. H. (2022). My tutor is an AI: The effects of involvement and tutor type on perceived quality, perceived credibility, and use intention. In H. Degen & S. Ntoa (Eds.), *Artificial intelligence in HCI* (pp. 232–244). Springer.

Demartini, G., Mizzaro, S., & Spina, D. (2020). Human-in-the-loop artificial intelligence for fighting online misinformation: Challenges and opportunities. *IEEE Data Engineering Bulletin, 43*(3), 65–74.

Dicicco-Bloom, B., & Crabtree, B. F. (2006). The qualitative research interview. *Medical Education, 40*(4), 314–321.

Dwivedi, Y. K., Kshetri, N., Hughes, L., et al. (2023). "So what if ChatGPT wrote it?" Multidisciplinary perspectives on generative AI. *International Journal of Information Management, 71*, 102642.

Edgerly, S., Mourão, R. R., Thorson, E., & Tham, S. M. (2020). When do audiences verify? *Journalism & Mass Communication Quarterly, 97*(1), 52–71.

Endacott, C. G. (2024). Enacting machine agency in AI communication technologies. *Journal of Computer-Mediated Communication, 29*(4), zmae011.

Faruk, L. I. D., Rohan, R., Ninrutsirikun, U., et al. (2023). University students' acceptance of generative AI. In *Proceedings of the international conference on advances in information technology* (pp. 1–8). ACM.

Fui-Hoon Nah, F., Zheng, R., Cai, J., et al. (2023). Generative AI and ChatGPT: Applications and challenges. *Journal of Information Technology Case and Application Research, 25*(3), 277–304.

Gibson, J. J. (1977). The theory of affordances.

Glaser, B. G., & Strauss, A. L. (2017). *The discovery of grounded theory: Strategies for qualitative research*. Routledge.

Goh, D. H. (2024). Media and strategies for deepfake identification. *Journal of the Association for Information Science and Technology, 75*(6), 643–654.

Grady, M. P. (1998). *Qualitative and action research: A practitioner handbook*. Phi Delta Kappa Educational Foundation.

Helmus, T. C. (2022). *Artificial intelligence, deepfakes, and disinformation*. RAND Corporation.

Himma-Kadakas, M., & Ojamets, I. (2022). Debunking false information. *Digital Journalism, 10*(5), 866–887.

Hong, J. W., Peng, Q., & Williams, D. (2021). AI music and expectancy violation theory. *New Media & Society, 23*(7), 1920–1935.

Kang, H., & Lou, C. (2022). AI agency vs. human agency. *Journal of Computer-Mediated Communication, 27*(5), zmac014.

Khemani, B., Patil, S., Kotecha, K., et al. (2024). Detecting health misinformation. *MethodsX, 12*, 102737.

Koh, Y. J., & Sundar, S. S. (2010). Heuristic versus systematic processing. *Human Communication Research, 36*(2), 103–124.

Krueger, R. A., & Casey, M. A. (2000). *Focus groups: A practical guide for applied research*. SAGE.

Larsson, S., & Heintz, F. (2020). Transparency in artificial intelligence. *Internet Policy Review*.

Liew, T. W., Tan, S. M., Yoo, N. E., et al. (2023). AI chatbots in health communication. *Computers in Human Behavior Reports, 11*, 100323.

Lincoln, Y. S., & Guba, E. G. (1985). *Naturalistic inquiry*. SAGE.

Liu, B. (2021). In AI we trust? *Journal of Computer-Mediated Communication, 26*(6), 384–402.

Metzger, M. J. (2007). Credibility on the Web. *Journal of the American Society for Information Science and Technology, 58*(13), 2078–2091.

Metzger, M. J., & Flanagin, A. J. (2015). Psychological approaches to credibility assessment. In S. S. Sundar (Ed.), *The handbook of the psychology of communication technology* (pp. 445–466). Wiley.

Metzger, M. J., Flanagin, A. J., & Medders, R. B. (2010). Credibility evaluation online. *Journal of Communication, 60*(3), 413–439.

Monteith, S., Glenn, T., Geddes, J. R., et al. (2024). Artificial intelligence and misinformation. *The British Journal of Psychiatry, 224*(2), 33–35.

Ooi, K. B., Tan, G. W. H., Al-Emran, M., et al. (2023). Generative AI across disciplines. *Journal of Computer Information Systems*.

Park, H. E. (2024). Generative AI in digitalization. *Psychology & Marketing, 41*, 2924–2941.

Patton, M. Q. (2014). *Qualitative research & evaluation methods*. SAGE.

Pundir, V., Devi, E. B., & Nath, V. (2021). Fake news sharing on social media. *Management Research Review, 44*(8), 1108–1138.

Qian, S., Shen, C., & Zhang, J. (2023). Fighting misinformation. *Journal of Computer-Mediated Communication, 28*(1), zmac024.

Rai, A. (2020). Explainable AI. *Journal of the Academy of Marketing Science, 48*, 137–141.

Rubin, H. J., & Rubin, I. S. (2011). *Qualitative interviewing*. SAGE.

Saunders, B., Sim, J., Kingstone, T., et al. (2018). Saturation in qualitative research. *Quality & Quantity, 52*(4), 1893–1907.

Shin, D. (2022). Perception of humanness in AI journalism. *New Media & Society, 24*(12), 2680–2704.

Shin, D. (2023). *Algorithms, humans, and interactions*. CRC Press.

Shin, D. (2024). *Artificial misinformation*. Palgrave Macmillan.

Shin, D., & Park, Y. J. (2019). Algorithmic affordance. *Computers in Human Behavior, 98*, 277–284.

Sun, Y. (2022). COVID-19 misinformation verification. *Science Communication, 44*(3), 261–291.

Sun, Y., Chen, J., & Sundar, S. S. (2024). Chatbot ads with human touch. *Journal of Business Research, 172*, 114403.

Sundar, S. S. (2020). Rise of machine agency. *Journal of Computer-Mediated Communication, 25*(1), 74–88.

Tandoc, E. C., Ling, R., Westlund, O., et al. (2018). Authentication in fake news. *New Media & Society, 20*(8), 2745–2763.

Thomson, T. J., Angus, D., Dootson, P., et al. (2022). Visual misinformation verification. *Journalism Practice, 16*(5), 938–962.

Torres, R., Gerhart, N., & Negahban, A. (2018). Credibility on social media. *ACM SIGMIS Database, 49*(3), 78–97.

Tracy, S. J. (2020). *Qualitative research methods*. Wiley-Blackwell.

Tsang, S. J. (2022). COVID-19 misinformation tactics. *Online Media and Global Communication, 1*(3), 469–496.

Waruwu, B. K., Tandoc, E. C., Duffy, A., et al. (2021). Social authentication. *New Media & Society, 23*(9), 2516–2533.

Wathen, C. N., & Burkell, J. (2002). Credibility on the Web. *Journal of the American Society for Information Science and Technology, 53*(2), 134–144.

Wiles, R. (2012). *What are qualitative research ethics?* Bloomsbury Academic.

Xu, D., Fan, S., & Kankanhalli, M. (2023). Combating misinformation. In *Proceedings of the ACM international conference on multimedia* (pp. 9291–9298).

Zaichkowsky, J. L. (1994). Personal involvement inventory. *Journal of Advertising, 23*(4), 59–79.

Zhou, Y., & Shen, L. (2022). Confirmation bias and misinformation. *Communication Research, 49*(4), 500–523.