Article │ Received 1 February 2026; Revised 9 March 2026; Accepted 17 March 2026; Published 19 March 2026
https://doi.org/10.55092/let20260003

# Artificial integrity: social performativity of honesty in the age of generative AI

**Alma Espartinez**[1,2]

[1] Theology/Philosophy Department, De La Salle-College of Saint Benilde, Manila, Philippines
[2] Philosophy Department, Providence College, Providence, USA

E-mail: alma.espartinez@benilde.edu.ph.

## Highlights:

- Students use AI yet claim honesty.
- Artificial integrity is technologically scaffolded self-deception.
- Institutional ambiguity enables rationalization space.
- Process transparency collapses performative authorship.
- Reform requires assessment redesign and not detection.

**Abstract:** This article analyzes the paradoxical phenomenon in which students extensively utilize generative AI for academic work while sincerely maintaining that their submissions are honest and original. Beyond simple confusion or concealment, it introduces artificial integrity: a techno-ethical dilemma arising from technologically scaffolded knowing self-deception. Drawing from dramaturgical analysis, narrative identity theory, and recent empirical research, a framework is developed that reveals how integrity is socially performed and stabilized within ambiguous institutional ecologies. The analysis demonstrates that students, while retaining awareness of AI's core intellectual labor, sustain credible honesty claims through epistemic layering, manifesting in strategic disclosure, resistance to transparency, and persistent anxiety. This condition is co-produced by institutional designs that prioritize polished outputs over visible process, creating a rationalization space where traditional legal-ethical frameworks for authorship and accountability break down. Rather than policing AI use, this article argues institutions must develop clear, legally sound AI-use policies and redesign assessment to mandate transparency, through methods such as process portfolios, reflective annotations, and structured disclosure protocols, thereby resetting the academic stage to reward visible cognition over performative authorship.

**Keywords:** artificial integrity; academic integrity; generative AI; performativity; dramaturgy; higher education; knowing self-deception; epistemic layering

## 1. Introduction

A striking puzzle emerges: students acknowledge that generative AI (GenAI) produced most of their essay, yet insist the work remains honest, authentic, and fundamentally theirs. Students openly acknowledge reliance on GenAI while defending their actions as "collaboration," "curation," or "assistance" [1,2]. They disclose to lenient instructors, conceal from strict ones, and argue rather than apologize when challenged [3]. Notably, they demonstrate accurate knowledge of institutional policies and can identify violations in others' behavior [4]. The puzzle is how students sustain coherent moral self-understanding when core intellectual labor has been machine-performed.

I name this condition artificial integrity*:* a technologically scaffolded knowing self-deception that allows students to feel honest. Across studies, students calibrate their disclosures to match instructor leniency [5,6]. Asked to quantify AI contributions, they resist [7,8], justifying AI use as curation while acknowledging its generative role [7]. Yet, beneath stated confidence, anxiety persists despite believing they have followed the rules [8,9]. Recent surveys capture two conflicting trends: widespread adoption and pervasive underreporting [10]. Up to 90% of students believe their peers use AI, while only 60% admit to using it themselves—a gap reflecting social desirability bias and lingering stigma [8]. Meanwhile, disclosure policies remain fragmented. Journals take different positions: some mandate full transparency, others treat AI assistance as incidental and not worth mentioning [11]. This contradiction between high usage and low disclosure, set against ambiguous institutional guidance, creates the precise conditions for artificial integrity to emerge.

### *1.1. Artificial integrity: what it is and is not*

Artificial integrity describes a specific psychological and social condition: knowing self-deception—holding contradictory knowledge without being forced to resolve it. Knowing self-deception occurs when students foreground self-serving justifications—"collaboration" or "guidance"—while retaining, in the background, accessible knowledge that AI generated the major content. That background knowledge does not disappear; it lingers, shaping behavior, especially when accountability looms.

Drawing on evolutionary and psychological accounts of self-deception, artificial integrity captures a state in which people neither fully believe their rationalizations nor abandon them [12,13]. This condition differs from three related phenomena frequently invoked in discussions of academic misconduct:

- Not genuine confusion, where contradictory knowledge is absent or unresolved. Students demonstrate accurate knowledge of policy and can reliably identify violations when evaluating others' work [4].
- Not simple concealment, where wrongdoing is privately acknowledged but strategically hidden without a self-narrative to justify it [14].
- Not motivated belief formation, where selective exposure and biased reasoning produce a sincere revision of belief that resolves the contradiction [15,16].

*Knowing self-deception* holds the contradiction rather than resolving it. A parallel track of awareness persists, guiding behavior when accountability intensifies—surfacing in selective disclosure, resistance to transparency, and anxiety about detection even in the absence of any challenge. This is what epistemic layering describes: motivated interpretations foregrounded in low-stakes contexts, while contradictory knowledge remains accessible, actionable, and ready to shape behavior when evaluative pressure rises

or audiences shift. Students are not ignorant; they are operationally informed while narratively invested elsewhere. Together, knowing self-deception and epistemic layering explain how students sincerely embody integrity in some settings while their behavior—calibrated, opaque, anxious—betrays retained awareness of cognitive substitution in others. Artificial integrity is not a psychological anomaly. It is a patterned, socially intelligible condition emerging under specific institutional constraints: where policies are ambiguous, processes invisible, and assessment rewards the polished product over the visible labor behind it [17,18]. Students sustain honesty claims by managing rather than resolving a known contradiction—compartmentalizing rather than eliminating it, preserving self-concept while outsourcing the labor [19]. That is what this framework names—and what institutions must be redesigned to address.

### *1.2. How epistemic layering operates*

Integrity narratives are not held as stable propositions; they function as context-sensitive interpretive frames. In low-stakes or socially supportive settings, rationalizations are foregrounded. Students experience their AI-assisted work as legitimate, guided, and aligned with institutional values—interpretations that feel sincere and defensible, particularly when supported by grades or instructor approval. Under heightened accountability, however, background knowledge becomes salient. Students adjust disclosure, avoid transparency, and prepare defenses. The same student moves across multiple epistemic registers without resolving the underlying contradiction.

This layering produces a distinctive behavioral pattern. Students display accurate policy knowledge while selectively exempting themselves. They express confidence in their work while maintaining persistent anxiety about its evaluation. They defend their practices when challenged, yet resist articulating them in detail [9]. Students do not carry integrity as a fixed belief; they carry different versions of it into different rooms. These behaviors do not indicate hypocrisy or confusion; they indicate the retained operational relevance of contradictory knowledge [20]—awareness that cannot be fully admitted without destabilizing the narrative of authorship. These behavioral markers are formalized as testable propositions in Section 3.2.

### *1.3. How generative AI scaffolds knowing self-deception*

GenAI stabilizes knowing self-deception through three affordances that make rationalization unusually sustainable.

Anthropomorphism. Conversational interfaces invite students to treat AI as a dialogic partner, enabling "collaboration" narratives that students recognize as euphemistic yet deploy for socially acceptable framing. A student who prompts ChatGPT to develop an argument may later describe the exchange as "dialogue" or "thought partnership," even while recognizing that AI generated the core ideas [21]. In peer study groups, students frame AI use as "getting feedback" and "brainstorming together," a narration that feels natural and poses no challenge. They foreground their role in guiding the process while background awareness acknowledges that "a major chunk of the thinking had been done for me" [22]. This managed duality is what artificial integrity describes.

Process invisibility. Opacity enables "unmeasurable contribution" claims. Students could, if pressed, articulate what they contributed *versus* what the AI produced. Yet ambiguity shields their claims from falsification. A student may describe "iterative prompting" while declining to share prompt history,

preserving plausible deniability about the extent of cognitive outsourcing [23]. When asked how a thesis was developed, the same student offers generalities about "refinement" while omitting that ChatGPT generated the initial draft. They have not forgotten; they shift which knowledge guides behavior as the audience changes. Process invisibility makes such calibration possible: authorship can be performed because its backstage mechanics remain opaque [7].

Originality illusions. Novel outputs enable "not plagiarism" defenses that substitute textual originality for intellectual authorship. A student submits an entirely AI-generated essay that passes plagiarism detection and insists, "I did not copy anyone; the text is original" [24]. The defense reframes authorship as mere non-duplication. This produces the familiar dual awareness: foreground confidence in curatorial judgment alongside background doubt about whether they could reproduce the analysis under scrutiny. The output is new; the intellectual labor is displaced.

Research suggests that perceived peer behavior, more than policy knowledge, predicts academic misconduct by lowering perceived detection risk [25]. Artificial integrity extends this logic. Anthropomorphism, opacity, and novelty supply rationalization resources; peer normalization lowers the cost of deploying them. Within permissive institutional spaces, students retain awareness that their justifications are self-interested, evidenced by avoidance of contexts requiring explicit articulation of cognitive labor, declining to quantify AI contributions, refusing to share prompts, and hesitating under direct inquiry [6,10]. Disclosures shift accordingly: casual acknowledgment to peers who also use AI, strategic reframing to judgmental audiences, defensive elaboration when pressed. Qualitative studies report similar patterns of authorship claims accompanied by anxiety and guarded transparency [7,26].

GenAI does not erase contradictory knowledge. It scaffolds its sustainability. Students uphold integrity values while outsourcing cognitive labor because the contradiction is managed rather than resolved. The central question is not whether they know AI did the work—*they do*—but how they preserve moral self-concept despite that knowledge.

### *1.4. Analytical rather than normative approach*

This article adopts an analytical rather than a normative stance, seeking to understand how students use GenAI extensively while claiming honesty without apparent irony [20]. This analytical focus does not imply moral neutrality. On the contrary, the framework highlights the co-production of artificial integrity, revealing how institutional ecologies (through ambiguous policies, product-focused assessments, and inconsistent enforcement) actively support the rationalization space that students inhabit and exploit. The focus shifts accordingly from individual culpability to systemic design. How institutions should respond—whether to dismantle, accommodate, or redefine artificial integrity—depends on contested values regarding learning, assessment, and the purpose of higher education. The article does not put students on trial. It examines the system that makes their behavior rational.

### *1.5. Scope and boundary conditions*

The framework of artificial integrity applies most clearly under specific institutional conditions. It does not claim to explain all student conduct, but rather a patterned response that emerges when certain features of higher education converge.

Artificial integrity is likely to arise when:

- Assessment is product-focused, centered on polished outputs rather than visible intellectual work. Opacity of process creates space for students to perform authorship without disclosing how they got there.
- Policies on AI use are ambiguous**,** contradictory, or unevenly enforced. Vague guidance like "use AI responsibly," paired with inconsistent instructor practices, creates interpretive flexibility, allowing students to craft defensible justifications without clearly violating stated rules.
- Disciplinary norms do not require visible reasoning**.** Artificial integrity takes hold most easily where core competencies do not demand real-time prose generation, and where intellectual struggle itself is not assessed. Large undergraduate courses that rely on standardized writing tasks are especially susceptible, reinforced by structural pressures toward efficiency and scalability [27].

Where these conditions are absent, the framework loses explanatory force. In process-transparent settings, such as laboratory notebooks, studio critiques, live coding, and oral examinations, the dramaturgical divide collapses. Where apprenticeship-based or mastery-oriented models prevail, intellectual development is monitored continuously, leaving little room for strategic opacity. Likewise, institutions with clear, coherent, and consistently enforced AI policies shrink the rationalization space by narrowing the range of plausible justifications [6].

These boundary conditions do not universalize artificial integrity. They specify the institutional ecology in which it becomes a stable and intelligible response. The framework explains most where product opacity, policy ambiguity, and permissive disciplinary norms intersect. As those conditions diverge, alternative explanations—genuine confusion or sincere belief revision—become more plausible.

For information technology education research, this framework clarifies why better policies and detection tools alone will not solve AI-related integrity challenges. In legally constrained settings, such measures may even backfire, introducing due-process risks, evidentiary uncertainty, and privacy concerns without resolving the underlying tension between performance and authorship. The framework reframes academic integrity as something institutions produce through their own design choices—a shift that opens different questions for those evaluating AI-integrated pedagogies and governance structures.

## 2. The social machinery of sincerity: a performative framework

Integrity is not an inner virtue to be uncovered but a credible identity to be accomplished. It emerges where institutional signals, social narratives, and staged performances converge to authenticate a student's work as their own, even when a language model wrote it. In this performative framework, honesty functions first as a recognizable social identity, second as a story one tells about oneself, and only indirectly as an internal moral state. The question shifts from "Are they honest?" to "How is honesty made credible in this context?" Three interlocking pillars constitute this machinery.

- Dramaturgical staging [28]. AI is not merely a cheating tool. It becomes a dramaturgical resource. Students perform authorship for evaluative audiences, using polished outputs as props and conversational interfaces as scripts. They retain backstage awareness of who actually did the thinking, but the performance, if credible, is enough.
- Narrative construction [29]. Students weave AI use into their moral self-concept through stories of collaboration, efficient synthesis, or augmented inquiry. These narratives feel authentic because they draw on culturally valued scripts: productivity, innovation, and working smarter.

A student who tells themselves they "partnered" with AI to refine their ideas is not necessarily lying. They are making their actions coherent within a self they can live with.

- Institutional ecology. Ambiguous policies, product-focused assignments, and inconsistent enforcement do not merely fail to prevent artificial integrity—they actively supply the conditions that make it rational.

These three pillars reinforce one another (see Figure 1). Students describe current academic-integrity policies as "laughably naïve" [22] in the age of GenAI. This gap between institutional rules and technological reality creates a space for rationalization. Students can plausibly ask: if the system cannot detect AI use and offers no clear guidance on acceptable collaboration, is it truly unethical to use it? Ambiguity is not a loophole; it is a structural feature of an institutional ecology that prioritizes polished products over visible processes. Ambiguity gives students room to perform authorship and craft convincing self-narratives. When those performances are rewarded with grades or approval, their rationalizations gain plausibility. Their fluency increases. The system tightens around them [30]. This validation does not resolve the underlying contradiction of knowing self-deception; it simply makes it sustainable. In this light, artificial integrity is not a glitch in the system, but a structural outcome. The institutional environment rewards the appearance of authentic work while obscuring the labor behind it. Under such conditions, performance becomes sincerity's most viable substitute.

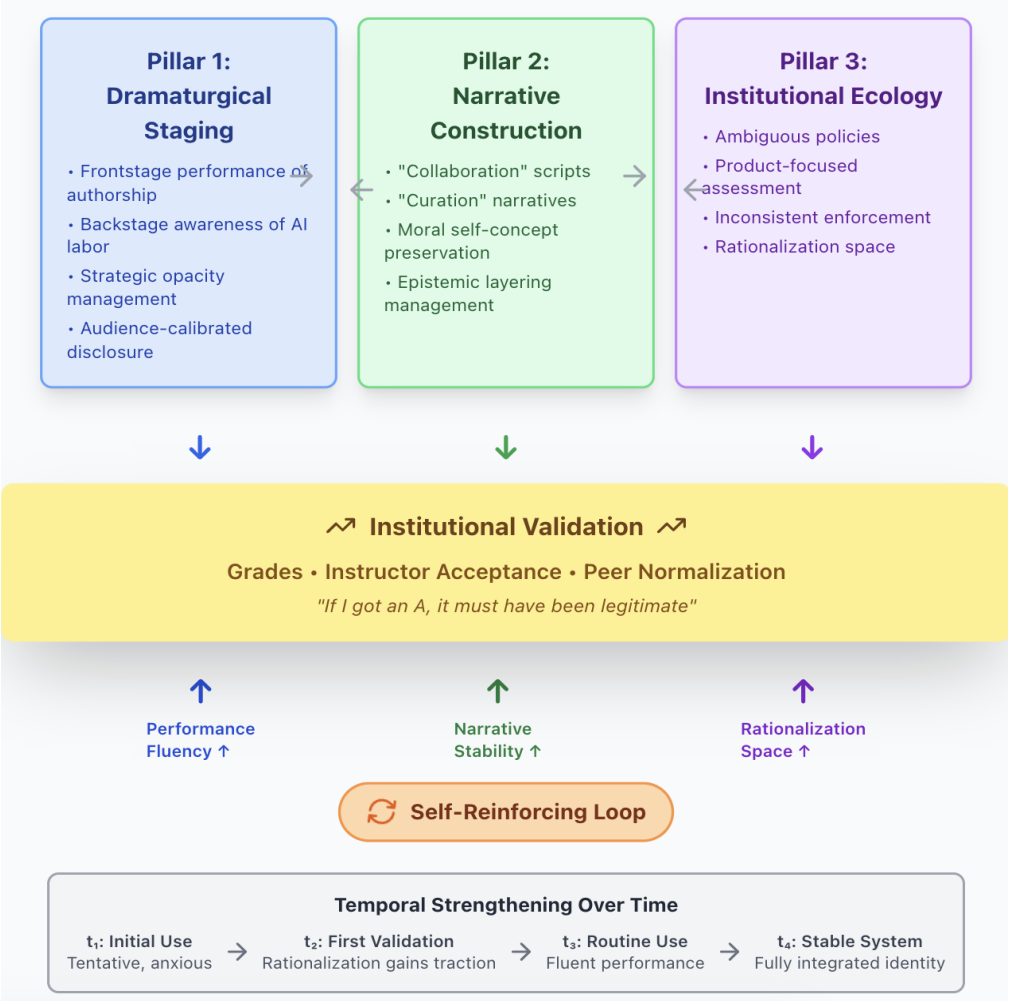


**Figure 1.** The pillars of artificial integrity as a performative framework.

The three pillars interact bidirectionally and feed into Institutional Validation—grades, acceptance, normalization—creating a self-reinforcing feedback loop. Each successful validation increases performance fluency, narrative stability, and rationalization space, strengthening all pillars over time. Students move from tentative use to fluent performance, but background awareness persists. Knowing self-deception does not resolve; it routinizes.

### *2.1. Pillar 1: dramaturgical staging: the performance of authorship*

From a Goffmanian perspective, GenAI functions less as a concealed violation than as a dramaturgical resource for staging credible authorship. Its affordances—anthropomorphic conversation, opaque processes, and novel outputs—become tools for the performance. The polished essay becomes a frontstage artifact; the AI's cognitive labor stays carefully backstage. Students know who did the thinking, but they manage that awareness purposefully. They reveal AI use to some instructors while concealing it from others, avoiding contexts that demand transparency. Each successful validation—a good grade or an approving comment—makes the performance more fluent. Terms like "curation" and "prompt engineering" become second nature.

A concrete illustration appears in Sivasubramaniam's [6] documentation of students who, when asked by instructors to describe their writing process in post-submission interviews, offered fluent and confident accounts of their own reasoning—accounts that became noticeably vague when questions turned to the specific role AI had played at each stage. The frontstage narrative had been rehearsed; the backstage mechanics had not been prepared for scrutiny. This is dramaturgical staging in operation: the same underlying action performed differently depending on the perceived risk of each audience, with backstage awareness managed rather than disclosed.

But fluency is not belief. Backstage awareness never disappears. It shows up in detection anxiety, in calibrated disclosure, and in the refusal to share a prompt history. This is performative sincerity: the student enacts the role of author with sufficient conviction to satisfy the front stage, while backstage awareness is managed rather than resolved. Empirical evidence confirms this dynamic. Gonsalves [7] documents a systematic pattern of non-compliance with AI contribution declarations: students who acknowledge using AI for assistance nonetheless resist or decline to specify the nature or extent of that contribution when directly asked. This pattern is precisely what dramaturgical staging predicts. The performance of authorship requires that its backstage mechanics remain opaque, not merely to the evaluator, but as an active accomplishment of disclosure management.

### *2.2. Pillar 2: narrative construction—the stories we tell ourselves*

If dramaturgical staging explains how students perform honesty, narrative construction explains how they make sense of doing it. Students are not just acting; they are authoring. Faced with AI-assisted work, they weave AI use into a coherent moral identity: the dialogic partner, the resourceful professional, the critical thinker. The story is not necessarily a lie; it is a cognitive and moral bridge. It protects the "I" from the contradiction of the work.

Externally, these stories manage impressions. Internally, they maintain moral coherence by converting potential guilt into justification, thereby making cognitive dissonance more manageable. Students do not merely rationalize after the fact; they narrate prospectively, constructing in advance the

story that will make AI use morally intelligible to themselves and others. Qualitative evidence supports this account. Espartinez [5] finds that, across Philippine higher education contexts, students consistently frame AI-assisted work as collaborative inquiry. They position themselves as the director of a process rather than recipient of an output—a narrative form that preserves authorial identity while substantially displacing cognitive labor. A sharper illustration comes from Bowen and Watson [3], who document a business student explaining to an instructor that she had "used AI the way a consultant uses a research assistant—I directed the inquiry, evaluated the outputs, and made the judgment calls." When the instructor asked her to reconstruct the core argument without her notes, she could not. The narrative had been fluent; the cognition had not. She had authored the story of authorship without authoring the argument itself. Ling, Kale, and Imas [8] further demonstrate that social desirability shapes which narrative students deploy. The story told to a peer who also uses AI differs systematically from the one offered to an instructor, consistent with epistemic layering's claim that foreground rationalizations are context-sensitive.

### *2.3. Pillar 3: institutional ecology—the stage itself*

Performances need a stage. Narratives need an audience. Artificial integrity needs an institutional ecology that quietly authorizes it. This ecology is built from ambiguous policies, product-focused assessments, inconsistent enforcement, and institutional values that prize efficiency over process. These are not loopholes but design features. They signal that what matters is the polished output, not the path taken. When detection seems unlikely, rationalization becomes the obvious response.

Universities thus co-produce the very integrity they claim to measure, providing the rationalization space that makes 'collaboration' narratives plausible and strategic opacity rational. But this space is not simply a policy vacuum. It is actively maintained by institutional choices: how assessments are designed, how consistently policies are enforced, and what signals are sent when AI use is tacitly tolerated in practice while formally prohibited in documents. The ecological character of this dynamic is borne out empirically. Theoharakis, Mylonopoulos, and Papadopoulou [31] find that the strongest predictor of AI-related academic misconduct is not individual policy knowledge but perceived peer behavior—students who believe their peers use AI extensively are significantly more likely to do so themselves, regardless of their own beliefs about the rules. This finding sits in tension with individually volitional accounts of academic integrity but is precisely what an ecological model predicts: when the institutional environment normalizes a behavior, individual ethical resistance requires exceptional effort, and rationalization fills the gap.

The concrete stakes of this ecology are visible in documented disciplinary cases. Gonsalves [7] reports that when students at one institution were formally asked to complete AI-use declarations, a substantial proportion left them blank or submitted vague responses — not because they had forgotten the requirement, but because the institution had offered no clear standard for adequate disclosure. The policy existed; the ecology made compliance optional. Students were not circumventing a rule so much as inhabiting an ambiguity the institution had designed into its own procedures. The institutional dimension of this dynamic is illustrated concretely by Resnik and Hosseini's [11] analysis of disclosure policy fragmentation across academic journals: some mandate full AI transparency, others treat AI assistance as incidental and unreportable, and many offer no guidance at all. Students navigating this landscape do not encounter a clear prohibition to circumvent; they encounter a genuinely ambiguous signal

about what honesty requires. The broader academic system does not merely fail to enforce integrity; it actively supplies the interpretive flexibility that makes rationalization not just possible but reasonable.

Artificial integrity, therefore, is not merely a student invention, nor is it a deterministic institutional imposition. It is a strategic adaptation co-produced within an ecology, making it a viable and often rational option. Students are active agents exploring and exploiting this space; the institution becomes a silent co-producer.

## 3. Research agenda: from theory to empirical investigation

The question is no longer how to measure dishonesty, but how to investigate the social machinery that sustains it.

### *3.1. Redirecting inquiry: from measurement to pillars*

Prevailing research on academic integrity often focuses on measurement—quantifying cheating prevalence, surveying attitudes toward AI, or testing detection tools. While valuable, this approach cannot capture the performative, narrative, and ecological processes that constitute artificial integrity. To examine self-deception in action, research must trace how rationalizations are deployed, how narratives are stabilized among peers, and how institutional cues are interpreted in real time.

The goal is not to verify whether students are honest, but to investigate how the social machinery of sincerity operates. This requires methods capable of capturing meaning-making in motion: ethnographic observation of disclosure practices, discourse analysis of student justifications, comparative case studies of departmental ecologies, and longitudinal interviews that track narrative development alongside institutional feedback. Such approaches shift the empirical focus from individual compliance to interactional, interpretive, and institutional dynamics. They move from asking "do students cheat?" to asking, "how do students sustain contradictory knowledge, and what social structures make this condition not only possible but durable?"

### *3.2. Testable propositions: operationalization and falsification*

The framework generates six falsifiable propositions that operationalize artificial integrity as knowing self-deception and distinguish it empirically from models of confusion or concealment. Table 1 summarizes each proposition and its core prediction, comparing it to alternatives.

To test these propositions, researchers must move beyond self-report surveys and employ methods sensitive to context and contradiction:

- Narrative Stability and Fluency can be measured through longitudinal interviews, analyzing coherence, elaboration, and consistency of justifications over time and across contexts. Decreased hedging might indicate fluency, but not necessarily belief.
- Strategic Disclosure and Opacity can be observed ethnographically or through experimental vignettes in which students choose what to disclose to different instructor profiles. Audience calibration is quantified as statistically significant differences in disclosure rates or narrative framing between strict and lenient conditions.
- Retained Awareness and Epistemic Layering are the most challenging to measure directly, as students may not be able to articulate them. Indirect markers include detection anxiety (self-report

or psychophysiological measures) despite integrity claims, resistance to transparency demands (e.g., refusal to share prompts in a safe, non-punitive research setting), and reaction-time or priming experiments that reveal automatic associations—comparing, for instance, 'my work' with 'AI-generated' *versus* 'collaboration'.

**Table 1.** Propositions of artificial integrity.

| Proposition | Core Claim | Key Research Question | Prediction (*vs.* Alternatives) |
|---|---|---|---|
| **P1** | Students with greater audience awareness deploy more sophisticated performative repertoires while managing, not resolving, contradictory knowledge. | How do performative repertoires (e.g., 'collaboration' narratives) vary by audience (strict *vs.* lenient instructors, peers)? | Students will show audience-specific narrative shifts and report managing dual awareness (not confusion or belief change). |
| **P2** | Narrative scripts aligned with disciplinary values (e.g., "efficient synthesis" in business) show greater stability, reinforced by institutional validation. | Which narratives are most stable, and how does disciplinary context shape their viability? | Narrative stability correlates with institutional validation but not with reduced strategic behavior (awareness persists). |
| **P3** | AI interface features (anthropomorphism, opacity, novelty) directly shape the rationalization resources students deploy. | How do different AI platforms (e.g., conversational *vs.* completion-based) enable different rationalization narratives? | Interface features shape rationalization content, but strategic disclosure and anxiety patterns persist across interfaces (awareness is retained). |
| **P4** | Ambiguous policies and product-focused assessments increase the prevalence and stability of artificial integrity by providing rationalization space. | Does policy clarity reduce artificial integrity, or merely shift rationalization strategies? | Ambiguity increases prevalence; clarity shifts rationalization but does not eliminate strategic behavior (confusion model falsified). |
| **P5** | Repeated successful performance increases narrative fluency and rationalization plausibility, but does not eliminate contradictory knowledge. | Does fluency correlate with reduced awareness, or merely with routinized management of retained knowledge? | Fluency increases, but strategic patterns (disclosure variance, anxiety) persist; validation increases plausibility, not belief. |
| **P6** | Validation (grades, instructor acceptance) creates a self-reinforcing loop: rewarded rationalizations are redeployed, strengthening performance fluency while awareness remains**.** | Do students abandon or adapt rationalizations after failed performances (low grades, challenge)? | Failed validation prompts narrative adaptation, not abandonment; successful validation strengthens redeployment. Awareness persists, evidenced by continued context-sensitive behavior. |

Note*:* Propositions are hierarchically organized: P1 and P2 test rationalization deployment, P3 and P4 examine contextual shaping, and P5 and P6 address temporal and systemic reinforcement. Predictions are designed to distinguish between artificial integrity and alternative models.

### *3.3. Falsification criteria and measurement challenges*

The theory of artificial integrity would be falsified by evidence demonstrating that the core pattern of knowing self-deception does not hold. Specifically:

- If students show no audience-calibration in disclosure or narrative (falsifying P1 and P6).
- If policy clarity and process-transparent assessment lead to the abandonment of strategic justifications and strategic behaviors, such as opacity (falsifying P4 and P5).
- If narrative fluency and repeated validation correlate with the disappearance of indirect markers of retained awareness (anxiety, opacity), suggesting a shift to sincere belief (falsifying P5).

A primary challenge is social-desirability bias [32] and the performative nature of the phenomenon itself—students are unlikely to directly admit to knowing self-deception in an interview. Therefore, research must rely on:

- Convergent multi-method evidence: Triangulating self-reports, observed behavior, and indirect experimental measures.
- Longitudinal designs: Tracking students across multiple assignments and policy contexts to see how narratives adapt rather than dissolve.
- "Safe disclosure" research protocols: Protocols that assure participants their responses are for research only and will not affect grades, eliciting more honest accounts.

Establishing this empirical foundation is a precondition for the institutional and pedagogical interventions Section 4 develops.

## 4. Implications for pedagogy and institutional practice

If artificial integrity is structurally produced, it can only be structurally dismantled. Understanding *artificial integrity* as performatively sustained through institutional validation reveals why interventions aimed at confusion or concealment fall short. If students knowingly deploy rationalizations validated by institutional signals, then rule clarification or detection threats may shift rather than eliminate the underlying dynamics.

### *4.1. Normative foundations of procedural integrity*

The preceding sections argue that artificial integrity is structurally produced and therefore requires a structural response. But why privilege process transparency over clearer rules or stronger detection? The case for procedural integrity rests on three converging normative traditions: educational ethics, learning theory, and procedural justice.

- Educational ethics. Biesta [33] distinguishes education from training or credentialing. Education requires an encounter with genuine uncertainty—the experience of not yet knowing. AI outsourcing does more than reduce effort; it bypasses this encounter, substituting probabilistic output for intellectual risk. Outcome-focused assessment is structurally blind to this substitution. It rewards the artifact of learning without requiring learning itself. Process transparency restores visibility to the struggle and makes it evaluable.
- Learning theory. Vygotsky's zone of proximal development [34] locates development in the space between independent capacity and guided performance. Crucially, it is the attempt, error, and revision within this space—not the polished product—that constitutes growth. Assessment aligned with learning must therefore capture process, not merely output. Procedural integrity is normatively superior not because it is harder to game, but because it evaluates what learning actually consists in.
- Procedural justice. Rawls's account of pure procedural justice [35] holds that fairness depends on the integrity of the process that generates outcomes. Under AI conditions, outcome-based assessment loses its epistemic warrant. Two students may submit indistinguishable work through radically different cognitive processes—one through sustained engagement, another through delegated cognition. If evaluation attends only to the product, it treats these as

equivalent and calls the result fair. But fairness detached from process becomes illusory. Generative AI exposes what was previously tolerable: product quality no longer reliably indicates intellectual labor. Procedural integrity restores coherence by anchoring evaluation in visible engagement rather than surface output.

Across these traditions, the conclusion converges: education requires encounter, development requires struggle, and fairness requires process visibility. Procedural integrity is the arrangement most consistent with all three. This normative claim does not mandate uniform implementation. Some disciplines already institutionalize process visibility—laboratory notebooks, studio critiques, clinical logs, live coding sessions. In such contexts, procedural integrity formalizes existing practice. More substantial redesign is required where assessment culture centers on polished end products, particularly in large-enrollment writing-intensive courses. Infrastructure, workload, and student perception present genuine constraints. Responsiveness to these differences does not weaken the normative case; it specifies its institutional translation.

In the context of artificial integrity, where institutions inadvertently stabilize knowing self-deception, procedural integrity addresses the problem at its root. It does not intensify surveillance or refine detection. It realigns evaluation with the conditions under which learning, fairness, and authenticity converge. Reconfiguring the stage rather than policing the performance makes intellectual labor visible again.

### *4.2. From rule clarification to pillar disruption*

Institutional responses to GenAI have largely followed a familiar script: clarifying policies, increasing surveillance, and strengthening sanctions. From the performative perspective developed here, these measures treat artificial integrity as a problem of confusion or concealment. However, if students are engaged in *knowing* self-deception, then clearer rules and better detectors do not dismantle the machinery; they merely provide new lines for the same performance.

Rule clarification alone often functions as a new script, not a solution. When policies specify permissible AI use without restructuring assessment practices, they provide language that can be incorporated into existing rationalizations (“I stayed within the rules”), while leaving intact the dramaturgical divide between frontstage performance and backstage awareness. Students adept at performative staging adapt quickly, aligning their narratives with policy terminology while retaining the background knowledge that authorship remains ambiguous.

Similarly, detection-oriented interventions may alter surface behavior without transforming the epistemic structure that sustains artificial integrity. Heightened monitoring can reduce overt AI use or increase anxiety, but it does not necessarily disrupt the narrative scripts that protect moral self-coherence. A student who refrains from using AI out of fear of sanction may still perceive the institution as unfair or out of touch. Reduced misconduct does not signal a resolved contradiction; it only reflects constrained performance.

A performative account, therefore, redirects attention from rule enforcement toward pillar disruption. Interventions matter to the extent that they alter the conditions under which performances are staged, narratives are sustained, and institutional signals are interpreted. For example, process portfolios, reflective annotations, or live oral defenses collapse the dramaturgical space by making cognitive labor visible. When assessments require visible engagement with the cognitive process, when narratives must be reconciled with observable traces of work, and when evaluative cues are stable across contexts, the rationalization

space on which artificial integrity depends shrinks. Integrity becomes less a matter of strategic alignment with institutional ambiguity and more a transparent correspondence between action, account, and evidence.

### *4.3. Reinterpreting "AI-resilient" assessment through the performative lens*

So-called "AI-resilient" assessments—process portfolios, reflective annotations, viva voce exams, and iterative drafting—are often pitched as technical fixes. But from the performative perspective developed here, they are more than safeguards. They are reconfigurations of the stage itself. They do not just make cheating harder; they reshape the very conditions under which honesty can be credibly performed. First, they collapse dramaturgical space. Process-focused assignments (annotated drafts, prompt histories, reasoning trails) strip away the opacity that artificial integrity needs in order to survive. When every cognitive move is documented, there is no hidden backstage. The instructor is not just evaluating a final product; they are witnessing the labor. In this transparent theater, euphemisms like "collaboration" or "curation" collapse under the weight of visible evidence. A student cannot credibly claim a partnership with AI when the record shows who actually did the work. The performance has nowhere to hide. Second, they constrain narrative flexibility. Reflective components compel students to craft coherent narratives of their own learning. However, there is a catch: the story must align with the documented evidence. A student cannot comfortably narrate a "collaborative inquiry" if their process journal shows only passive copy-pasting. The narrative is no longer a free-floating justification; it is accountable to a visible trail of work. This forces a confrontation between the convenient scripts of artificial integrity and the actual, often messy, record of engagement. The protective interpretation fails when it has to square itself with the facts. Finally, and most systemically, they redesign the institutional ecology. When an entire program adopts consistent, process-transparent assessment, it sends a new signal: what matters here is not the polished output, but the demonstration of learning. This changes the currency of success. Efficiency and novelty take a backseat to struggle, iteration, and visible growth. In such an ecology, the rationalizations that fuel artificial integrity lose their cultural resonance. The system no longer rewards the performance; it rewards the process behind it.

This is not about making assignments AI-proof—a goal always temporary and is now plainly lost. It is about redesigning the stage itself, so that the performance of honesty and honest performance become the same thing. In that system, integrity is not something you claim. It is something you show.

### *4.4. Shifting from individual ethics to institutional integrity*

The preceding subsections establish what institutions should stop doing. This one addresses what they must do instead. The performative framework ultimately demands a reflexive turn: away from policing student ethics and toward examining how institutions inadvertently engineer conditions for artificial integrity. The institutional ecology, through its policies, assessments, and reward signals, does not merely enable dishonest behavior; it actively scaffolds a form of sincerity that is strategically viable, narratively coherent, and pragmatically rational. When the system rewards polished output over visible process, ambiguity over clarity, and efficiency over struggle, it is not failing to detect dishonesty; it is producing a new kind of honesty.

The primary implication for pedagogical design, therefore, is institutional reflexivity. Educators and administrators must ask not only, "Are our students cheating?" but also, "How do our assignments,

policies, and grading practices create the rationalization space we lament?" An assignment demanding polished writing under time pressure, guided by a vague AI policy, and evaluated solely on the final product, is not a neutral task. It is the very platform on which artificial integrity thrives.

The goal, then, is not to outsmart students with better detectors but to redesign the academic stage so that strategic performance loses its power. This involves creating assessment routines where AI use is integrated visibly and accountably into the learning process. In such a system, AI transforms from a dramaturgical resource for staging competence into a legitimate tool within a documented cognitive workflow. Integrity stops being a matter of opinion. It becomes the visible thread connecting what someone says, what they actually do, and the facts that back it all up.

This shift moves the problem of integrity under AI out of the ethics seminar and into the architect's office. It moves the conversation beyond enforcement toward pedagogies of visible cognition: learning environments where the process of thinking is necessarily transparent, where narratives are grounded in observable traces of work, and where the institution's values are coherently expressed through its assessments. In such environments, artificial integrity becomes untenable. Honesty no longer requires a performance, because the work itself is the performance.

### *4.5. Legal and regulatory frontiers: from detection to governance*

Redesigning the stage, however, operates within legal constraints that cannot be ignored. A performative analysis of artificial integrity reveals pressing legal and regulatory dilemmas for educational institutions. AI-detection software, for instance, implicates student data privacy, particularly when student writing is processed as biometric or behavioral data, and raises concerns about algorithmic due process when accusations rely on opaque probabilistic outputs. Furthermore, liability for AI-assisted work remains ambiguous. If a professional licensure issue or degree revocation stems from an undetected AI-generated error, accountability is unclear: does it lie with student, institution, or AI provider?

These questions transcend technical challenges; they represent governance dilemmas at the intersection of education law, contract law (through student codes of conduct), and emerging AI regulations. A performative lens suggests that legalistic, detection-centric responses may amplify the dramaturgical stakes, driving artificial integrity practices further underground. A more sustainable framework would shift from surveillance and sanction toward transparency and contract, clearly defining the rights, responsibilities, and acceptable boundaries of human-AI collaboration within the educational agreement.

Existing legal precedents underscore these procedural concerns. Public universities must provide due process before imposing sanctions that affect a student's property interest in their education (Goss v. Lopez, 419 U.S. 565 (1975)). When AI-detection tools produce probabilistic outputs (e.g., "85% likely AI-generated"), using such evidence as definitive proof risks violating these principles. Courts have held that students facing suspension or expulsion are entitled to notice of charges and an opportunity to be heard (Dixon v. Alabama State Board of Education, 294 F.2d 150 (5th Cir. 1961)). While private institutions are not bound by constitutional due process, they must adhere to their own stated procedures and cannot act arbitrarily. The student code of conduct constitutes a contractual agreement, and courts apply contract principles of interpretation when reviewing disciplinary actions (Schaer v. Brandeis University, 432 Mass. 474 (2000)).

## 5. Conclusion

At stake is not whether students are honest, but how honesty becomes recognizable in an age of delegated cognition operating under unsettled legal and institutional definitions of authorship, accountability, and human-AI collaboration. The shift this article proposes is therefore conceptual: from honesty as an inner state to honesty as a social achievement. Integrity has never been only a matter of character; it has always been a matter of design.

The paradox of students using AI extensively while claiming honesty is not adequately explained by confusion or concealment. It is the signature of a new condition: artificial integrity—a technologically scaffolded form of knowing self-deception sustained by three interlocking mechanisms: dramaturgical staging, narrative construction, and institutional ecology. Together, these form a self-reinforcing circuit in which performances of authorship are calibrated for evaluative audiences, self-stories are revised to absorb AI assistance, and institutional ambiguity supplies the rationalization space that keeps the contradiction stable.

### *5.1. Alternative interpretations and a conditional stance*

Three interpretations of artificial integrity are possible: (a) students adapt rationally to contradictory institutional signals; (b) institutional ambiguity does not diminish students' moral responsibility; or (c) both are partly correct, and institutional redesign is preferable to either exculpation or punishment. This article adopts the third view while remaining analytically agnostic about ultimate moral judgment.

A fourth perspective deserves engagement: the claim that AI-assisted student work may constitute a legitimate epistemic practice that existing integrity frameworks fail to recognize. Scholars such as Bowen and Watson [3] and Hutson and Plate [4] argue that productive prompting, critical evaluation of outputs, and iterative refinement represent contemporary cognitive skills. On this view, the deficit lies not in students but in outdated assessment regimes.

This argument is not implausible, and the ecological analysis presented here is compatible with the claim that institutions have failed to update their evaluative frameworks. The divergence arises at the phenomenological level. The behavioral markers analyzed in this article—audience calibration, anxiety about detection, resistance to transparency, and the management rather than resolution of contradictory knowledge—are not consistent with settled epistemic confidence. They are consistent with knowing self-deception. If AI-assisted work were experienced as fully legitimate, we would expect open disclosure and narrative stability across contexts. Instead, we observe strategic variation.

The normative debate over the legitimacy of AI-assisted academic work remains unresolved. The artificial integrity framework does not settle that debate—it clarifies what is occurring beneath it. Institutional redesign is required regardless of how that debate ultimately resolves.

### *5.2. Analytic contribution*

This analysis makes three contributions.

- It identifies and explains a recurring empirical constellation—accurate policy knowledge, defensive justification, strategic audience management, and persistent anxiety—that confusion or concealment models cannot jointly account for.

- It specifies the social architecture that stabilizes this pattern: the interaction of performative staging, narrative construction, and institutional ecology.
- It demonstrates that artificial integrity is institutionally contingent. It thrives where policies are ambiguous, processes are opaque, and outcomes dominate evaluation. It recedes where process is visible, expectations clear, and intellectual labor demonstrable.

The challenge, therefore, is not merely to prevent dishonesty but to define what counts as credible honesty in a technologically distributed intellectual production. Artificial integrity is not primarily a moral defect in students. It is a structural feature of an ecology that rewards polished outputs while rendering cognitive labor invisible. Empirical evidence confirms that students underreport AI use precisely because the system makes concealment rational [10]. The path forward requires institutional courage: disclosure standards that are clear and consistently enforced, and assessment designs that reward visible cognition over polished product. Only when the stage is redesigned will artificial integrity cease to be the rational response to contemporary academic conditions.

## Declaration of Generative AI and AI-assisted Technologies

During the preparation of this manuscript, the author used GenAI tools solely for language refinement and copy-editing purposes. The core theoretical framework, conceptual contributions, and analytical claims originate entirely with the author. The author takes full responsibility for all content.

## Acknowledgments

No funding was received for this article.

## Conflicts of interest

The author declares no conflicts of interest.

## References

[1] Kasneci E, Seßler K, Küchemann S, Bannert M, Dementieva D, *et al.* ChatGPT for good? On opportunities and challenges of large language models for education. *Learn. Individ. Differ.* 2023, 103:102274.

[2] Perkins M. Academic integrity considerations of AI large language models in the post-pandemic era: ChatGPT and beyond. *J. Univ. Teach. Learn. Pract.* 2023, 20(2):1–24.

[3] Bowen JA, Watson CE. *Teaching with AI: a Practical Guide to a New Era of Human Learning*, 2nd ed. Baltimore: Johns Hopkins University Press, 2025.

[4] Hutson J, Plate D. *The Case Against Disclosure: Defending Creative Autonomy in the Age of AI.* Champaign: Common Ground Research Networks, 2025.

[5] Espartinez A. Bridging the educational divide with ChatGPT's integration in Philippine higher education: Q methodology and narrative inquiry studies. *JITE:IIP* 2025, 24:11.

[6] Sivasubramaniam S. Deterring academic integrity breaches: the roles of institutions, academics, and support services. *J. Sch. Publ.* 2025, 56(2):240–268.

[7] Gonsalves C. Addressing student non-compliance in AI use declarations: implications for academic integrity and assessment in higher education. *Assess. Eval. High. Educ.* 2025, 50(4):592–606.
[8] Ling Y, Kale A, Imas A. Underreporting of AI use: the role of social desirability bias. *SSRN* 2025, SSRN 5464215.
[9] Bjelobaba S, Waddington L, Perkins M, Folt ýnek T, Bhattacharyya S, *et al.* Maintaining research integrity in the age of GenAI: an analysis of ethical challenges and recommendations to researchers. *Int. J. Educ. Integr.* 2025, 21:18.
[10] Simon J. Generative AI, quadruple deception & trust. *Soc. Epistemol.* 2026, 40(1):101–115.
[11] Resnik DB, Hosseini M. Disclosing artificial intelligence use in scientific research and publication: When should disclosure be mandatory, optional, or unnecessary? *Account. Res.* 2026, 33(2).
[12] Von Hippel W, Trivers R. The evolution and psychology of self-deception. *Behav. Brain Sci.* 2011, 34(1):1–16.
[13] Trivers R. *The Folly of Fools: the Logic of Deceit and Self-deception in Human Life*. New York: Basic Books, 2011.
[14] Grigoryeva MS. A theory of concealment. *Theor. Soc.* 2024, 53:1321–1355.
[15] Kunda Z. Motivated inference: self-serving generation and evaluation of causal theories. *J. Pers. Soc. Psychol.* 1987, 53(4):636–647.
[16] Kunda Z. The case for motivated reasoning. *Psychol. Bull.* 1990, 108(3):480–498.
[17] Ateeq A, Alzoraiki M, Milhem M, Ateeq RA. Artificial intelligence in education: implications for academic integrity and the shift toward holistic assessment. *Front. Educ.* 2024, 9.
[18] Nwokocha GC, Morkel JDV. Navigating online learning and artificial intelligence: identifying and managing assessment risks in private higher education in South Africa. *IJEDE* 2025, 40(2):1–46.
[19] Ariely D. *The Honest Truth about Dishonesty: How We Lie to Everyone—Especially Ourselves*. New York: Harper Perennial, 2012.
[20] Duah JE, McGivern P. How generative artificial intelligence has blurred notions of authorial identity and academic norms in higher education, necessitating clear university usage policies. *Int. J. Inf. Learn. Technol.* 2024, 41(2):180–193.
[21] Epley N, Waytz A, Cacioppo JT. On seeing human: a three-factor theory of anthropomorphism. *Psychol. Rev.* 2007, 114(4):864–886.
[22] Terry OK. I'm a student. You have no idea how much we're using ChatGPT. The Chronicle of Higher Education. 2023. Available: https://www.chronicle.com/article/im-a-student-you-have-no-idea-how-much-were-using-chatgpt (accessed on 12 January 2026).
[23] Mazar N, Amir O, Ariely D. The dishonesty of honest people: a theory of self-concept maintenance. *J. Mark. Res.* 2008, 45(6):633–644.
[24] Bao A, Zeng Y. AI disclosure, moral shame, and the punishment of honesty. *Account. Res.* 2026, 33(3).
[25] Ehrmann T, Ludes LM, Reindl M. Does character matter when everyone cheats? Peer influence and environmental drivers of academic misconduct. *Kyklos* 2026, 79(1):112–128.
[26] Avery J, Abril PS, del Riego A. Attributing AI authorship: towards a system of icons for legal and ethical disclosure. *Nw. J. Tech. & Intell. Prop.* 2024, 22(1):1–54.
[27] Denker KJ. Student response systems and facilitating the large lecture basic communication course: assessing engagement and learning. *Commun. Teach.* 2013, 27(1):50–69.

[28] Goffman E. *The Presentation of Self in Everyday Life*. New York: Doubleday & Company, 1959.
[29] McAdams DP. The psychology of life stories. *Rev. Gen. Psychol.* 2001, 5(2):100–122.
[30] Sozon M, Alkharabsheh OHM, Fong PW, Chuan SB. Cheating and plagiarism in higher education institutions (HEIs): a literature review. *F1000 Research* 2024, 13:788.
[31] Theoharakis V, Mylonopoulos N, Papadopoulou K. AI's learning paradox: how business students' engagement with AI amplifies moral disengagement-driven misconduct. *Stud. High. Educ.* 2025, pp. 1–18.
[32] Krumpal I. Determinants of social desirability bias in sensitive surveys: a literature review. *Qual. Quant.* 2013, 47:2025–2047.
[33] Biesta GJJ. *Good Education in an Age of Measurement: Ethics, Politics, Democracy*. 2015. Available: https://api.taylorfrancis.com/content/books/mono/download?identifierName=doi&identifierValue=10.4324/9781315634319&type=googlepdf (accessed on 24 February 2026).
[34] Vygotsky LS, Cole M. *Mind in Society: Development of Higher Psychological Processes*. Cambridge: Harvard University Press, 1978.
[35] Rawls J. *A Theory of Justice: Original Edition*. Cambridge: Harvard University Press, 1971.