



ARTIFICIAL INTEGRITY AS KNOWING SELF-DECEPTION: COGNITIVE MECHANISM AND LEGAL LIABILITY IN GENERATIVE AI

Alma S. Espartinez

De La Salle – College of Saint Benilde,
Manila, Philippines

alma.espartinez@benilde.edu.ph

and

Providence College, Providence,
Rhode Island, USA

ABSTRACT

Aim/Purpose	This paper argues that while academic informing failures have traditionally been understood as either concealment or confusion, Generative AI (GenAI) design has introduced a third mode – one that current detection tools, integrity policies, and legal frameworks may be ill-suited to address.
Background	Generative AI destabilizes this concealment–confusion binary by enabling students to retain accurate knowledge of cognitive delegation while maintaining a morally acceptable self-understanding as authors. Disclosure consequently becomes a threat to identity rather than merely an obligation, creating a structural barrier to accurate client–agent informing.
Methodology	The study synthesizes informing science, cognitive psychology (knowing self-deception, epistemic layering), and legal doctrine (foreseeability, due process) to diagnose how GenAI design disrupts the informing chain and to derive normative implications for developers, institutions, and regulators.
Contribution	This paper (1) identifies knowing self-deception as the mechanism that inverts the informing relationship; (2) shows how three AI design affordances externalize and stabilize this mechanism; (3) advances five falsifiable propositions for restoring informing integrity; and (4) argues that these effects are arguably foreseeable, which – if sustained – would ground design-level responsibility and highlight apparent structural gaps in existing regulatory frameworks.
Findings	Artificial integrity is theorized as a structurally durable informing failure sustained through epistemic layering, scaffolded by AI design, and reinforced by

Accepting Editor Akhil Prakash | Received: March 31, 2026 | Revised: May 28, 2026 |

Accepted: June 6, 2026.

Cite as: Espartinez, A. S. (2026). Artificial integrity as knowing self-deception: Cognitive mechanism and legal liability in generative AI. *Informing Science: The International Journal of an Emerging Transdiscipline*, 29, Article 11.

<https://doi.org/10.28945/5808>

(CC BY-NC 4.0) This article is licensed to you under a [Creative Commons Attribution-NonCommercial 4.0 International License](https://creativecommons.org/licenses/by-nc/4.0/). When you copy and redistribute this paper in full or in part, you need to provide proper attribution to it to ensure that others can later locate this work (and to ensure that others do not accuse you of plagiarism). You may (and we encourage you to) adapt, remix, transform, and build upon the material for any non-commercial purposes. This license does not permit you to use this material for commercial purposes.

	institutional ambiguity. Interventions underperform because they focus on compliance rather than on repairing the informing channel. A legally adequate response requires recognizing the client’s right to accurate information about cognitive labor, distributing responsibility across students, institutions, and developers, and redesigning systems to collapse rationalization space.
Recommendations for Practitioners	Implement transparency requirements (process logs, contribution statements) that make the informing act unavoidable, and process-focused assessment (draft histories, oral defenses) that shifts the client-agent interaction from outcome evaluation to process verification.
Recommendations for Researchers	Test the five falsifiable propositions. Examine how different institutional designs affect the stability of the informing relationship, and how courts and regulators respond to evidentiary gaps in the client-agent chain.
Impact on Society	Reframes AI misconduct as a system-design problem that undermines the fundamental informing contract between learners and credentialing bodies. Opens avenues for regulatory accountability beyond individual discipline, with implications for professional and civic contexts in which accurate accounts of cognitive labor are consequential.
Future Research	Comparative effectiveness of transparency versus detection interventions; meta-cognitive training conditions that restore accurate self-informing; judicial and regulatory responses to informing failures; extension of knowing self-deception to other AI-mediated domains.
Keywords	artificial integrity; knowing self-deception; informing failure; epistemic layering; developer responsibility; academic integrity law; GenAI; informing science

INTRODUCTION

Informing science frames education as a system in which clients – institutions, professions, and society at large – depend on agents to perform intellectual labor and accurately report its origins (Cohen, 1999, 2009; Gill et al., 2018). This client-agent relationship is the epistemological foundation of credentialing: when the informing chain functions properly, clients receive accurate signals about what cognitive work was done, by whom, and to what standard. Academic integrity frameworks have long served as the mechanism for maintaining this chain, resting on a binary assumption that breakdowns occur either through willful misinforming or informational deficit – a binary that generative AI (GenAI) has made obsolete. This paper establishes this claim through three behavioral anomalies that neither existing framework can explain.

The condition this paper examines – *artificial integrity* – is one in which students retain accessible knowledge that AI has performed substantial intellectual work, yet sustain a morally acceptable self-understanding as authors (Espartinez, 2026a). Such students are neither concealing nor confused; they have become unreliable informants to themselves and, consequently, to every client that depends on their honest accounts. This is a structural failure of the informing system, one with cognitive, institutional, and legal dimensions. The paper asks how GenAI design may systematically produce this failure, and what responsibilities it imposes on the developers, institutions, and regulators that constitute the broader informing infrastructure of higher education.

The paper advances a specific applied psychology account of why this informing failure is so enduring. The core mechanism is knowing self-deception: a stable cognitive state in which contradictory knowledge is neither eliminated nor suppressed but managed through motivated moral reinterpretation. GenAI does not merely enable such self-deception; it stabilizes and externalizes it through three design affordances – anthropomorphic interaction, process invisibility, and originality illusions – that

embed rationalization directly into the tool, making accurate informing psychologically costly to sustain. The informing science tradition holds that informing quality depends not only on agents' intentions but also on the cognitive, institutional, and technological conditions in which informing occurs (Cohen, 1999, 2009; Nahl & Bilal, 2007). This paper applies that insight to show that the degradation described here is consistent with, and arguably a foreseeable consequence of, deliberate design choices. This study, therefore, extends beyond diagnosis to a normative claim: that developers who deploy systems whose affordances foreseeably compromise the client-agent informing relationship in educational settings may bear responsibility for those effects, and that current regulatory frameworks appear structurally unprepared to assign it.

This paper is one of two companion works that develop the artificial integrity framework. The earlier study (Espartinez, 2026a) examines the sociological architecture: how the informing relationship is performed, narratively constructed, and co-produced within institutional ecologies. The present paper specifies the internal cognitive mechanism that makes inaccurate self-informing psychologically sustainable and examines how legal frameworks respond to design choices that render this failure structurally durable.

The paper proceeds as follows. It first argues that current detection and policy frameworks misdiagnose the informing failure they seek to address – targeting concealment or confusion. At the same time, the actual mechanism is managed contradiction, evidenced by three empirical anomalies: students who endorse integrity norms yet use AI without disclosure (integrity-behavior paradox), systematic variation in disclosure across contexts that reflects identity management rather than understanding (disclosure variance paradox), and persistent avoidance of declaration requirements despite policy awareness (declaration avoidance paradox). It then develops the theoretical framework and advances five falsifiable propositions to distinguish artificial integrity from competing accounts and test targeted interventions. Finally, it examines the legal, regulatory, and design implications: developer responsibility grounded in foreseeability, the categorical inadequacy of current regulatory frameworks, institutional accountability for ambiguity that sustains known informing failures, and the design interventions most likely to collapse rationalization space.

WHY CURRENT FRAMEWORKS FAIL: THE EMPIRICAL CASE FOR MISDIAGNOSIS

Current institutional responses to AI-related academic misconduct rest on a systematic misdiagnosis of the informing failure they seek to address. Detection technologies and policy interventions presuppose that the informing chain breaks because agents either misunderstand what accurate informing requires (informational deficit) or deliberately withhold accurate information (willful misinforming) (Bretag, 2016; Dawson, 2021). A growing body of empirical evidence challenges both assumptions and points toward a third, distinct failure mode that neither framework can address (Espartinez, 2026a).

Detection-based approaches fail on two levels. Technically, AI detection tools remain unreliable, with modest accuracy and significant false positives, preventing institutions from trusting the tools' reports (Weber-Wulff et al., 2023). Psychologically, the approach assumes that the risk of being caught primarily deters misconduct. Behavioral ethics research directly challenges this: dishonest behavior is often constrained less by the probability of detection than by the need to maintain a positive self-concept (Bandura, 1999). Where behavior is governed by self-understanding, intensifying detection does not compel accurate informing; it compounds the mechanisms that already make it difficult.

Policy clarification performs no better. Awareness of academic integrity rules correlates only weakly with compliance (Alsharefeen & Al Sayari, 2025; Dawson, 2021): students who accurately know the policies and can identify violations in others' work still engage in undisclosed AI use. The informing failure is not a deficit in the agent's knowledge. More damaging, clearer policy language may provide

additional vocabulary for rationalization rather than constrain it: the agent reinterprets new policy terms as resources for self-justification rather than constraints on behavior (Espartinez, 2025, 2026b).

Mandatory disclosure requirements expose the same pattern. Students consistently underreport AI use despite understanding what disclosure requires (Gonsalves, 2025; Resnik & Hosseini, 2026). The failure is not one of capacity – agents know what to report – but of willingness: disclosure makes the extent of AI involvement explicit, constraining the ambiguity that informal justifications depend on. Students withhold information not because they misunderstand the requirement but because meeting it would destabilize the self-understanding they are motivated to preserve.

The sharpest evidence of misdiagnosis lies in the values-behavior gap. Students who strongly endorse academic integrity norms nonetheless report substantial undisclosed AI use (Cotton et al., 2024; Rudolph et al., 2023) – not because they reject institutional values, but because they simultaneously uphold those values while also engaging in non-disclosure. Survey data confirm this selective self-presentation: students systematically underreport their own AI use relative to their estimates of peers' use (Espartinez, 2026a), a pattern shaped by identity management rather than accuracy. Integrity beliefs, further, do not reliably predict disclosure behavior (Bao & Zeng, 2026; Bobier et al., 2025), confirming that the informing failure operates independently of the agent's stated commitments.

These patterns converge on a single point: agents possess the knowledge required for accurate informing, yet do not act on it. The breakdown is not at the point of knowledge they know, nor at the point of deliberate concealment; *they do not privately acknowledge wrongdoing*. It occurs where knowledge must be integrated into a self-understanding capable of sustaining truthful informing. Neither the confusion nor the concealment frameworks can identify that failure, because neither recognizes it as a failure mode.

THE THEORETICAL GAP

Existing frameworks account for parts of this picture but not its full shape. Each captures a recognizable element of AI-assisted misconduct while leaving the defining feature – the stable coexistence of contradictory knowledge – unaddressed. I propose artificial integrity as a conceptual framework that explains what these accounts cannot jointly account for: how agents retain accurate knowledge of their own cognitive delegation while sustaining a morally acceptable self-understanding, and why this condition proves durable across time and institutional contexts.

Moral disengagement theory (Bandura, 2018) and rationalization accounts (Tsang, 2002) both identify mechanisms – euphemistic relabeling, diffusion of responsibility, post hoc justification – that are visible when students describe AI use as “collaboration” or attribute authorship to prompting skill. Their explanatory purchase is real but partial: both assume wrongdoing is first recognized and then reframed. The evidence here suggests justificatory narratives are constructed in ways that pre-empt recognition altogether, operating at the level of self-understanding before the informing obligation is ever constituted.

Motivated reasoning (Kunda, 1990) captures how goal-directed cognition selectively processes evidence to reach desired conclusions. It explains why students who want to believe their work is legitimate find reasons to do so. What it does not explain is the persistence of contradictory knowledge alongside the motivated conclusion – the anxiety, the calibrated disclosure, the avoidance of specificity – which are precisely the behaviors inconsistent with successful motivated belief revision. Where motivated reasoning resolves the tension, artificial integrity sustains it.

Impression management theory (Goffman, 1959; Leary & Kowalski, 1990) accounts for audience-calibrated behavior – why students disclose to lenient instructors and conceal from strict ones. It correctly identifies the social dimension of disclosure as performance. Its limitation is that it treats this as purely strategic surface behavior, leaving the internal cognitive structure – the simultaneous maintenance of contradictory knowledge – theoretically unaddressed. Impression management explains the performance; it does not explain what the performer knows backstage.

Cognitive dissonance theory (Festinger, 1957) predicts pressure toward the resolution of conflicting beliefs, but here the tension persists stably over time – a persistence that is itself the informing failure. Self-concept maintenance theory (Bandura, 1999) fits better. Still, it is typically applied to marginal deviations. It does not account for the scale and durability of AI-assisted substitution, nor for the technological environment’s role in externalizing and stabilizing the rationalization process. Together, these frameworks approach, but do not reach, the mechanism proposed in this paper.

Three questions remain unanswered across all these frameworks: how individuals maintain access to contradictory knowledge without resolving it; why AI-mediated environments stabilize this condition more effectively than prior contexts; and why detection- and clarification-based interventions consistently fail to restore the informing relationship. Artificial integrity addresses all three by locating the mechanism in how contradictory knowledge is managed at the level of self-understanding, and how that management is structurally supported by AI design.

THEORETICAL FRAMEWORK: ARTIFICIAL INTEGRITY AS KNOWING SELF-DECEPTION – A PROPOSED ACCOUNT

This section develops the theoretical core: the cognitive mechanism that makes artificial integrity psychologically sustainable, and the design features that make it structurally durable.

THE CORE DYNAMIC: EPISTEMIC LAYERING

Artificial integrity is best understood as a form of knowing self-deception sustained through epistemic layering (Butterworth et al., 2022; Mele, 2000). Disfavored knowledge remains dispositionally accessible but is overridden by motivated moral reinterpretation, fundamentally disrupting the agent’s capacity to serve as a reliable informant to both self and client. “Knowing” here does not imply continuous conscious awareness or explicit acknowledgment of wrongdoing; it refers to the continued accessibility of contradictory knowledge, evidenced by strategic behavior that anticipates potential challenge. This distinguishes artificial integrity from both concealment and confusion.

The distinction from concealment requires a precise statement, as the two conditions can appear superficially similar. In concealment-based misconduct, the agent privately acknowledges wrongdoing (the self-understanding is already settled), and the informing failure is a deliberate act of withholding (Grigoryeva, 2024). The agent performs the cognitive work against a stable private acknowledgment; the agent knows – yet decides not to disclose – and manages that decision. In artificial integrity, no such settled acknowledgment exists. Institutional ambiguity, the conversational design of the AI interface, and the availability of culturally legitimated scripts – “collaboration,” “curation,” “prompt engineering” – perform the epistemic work that would otherwise require deliberate internal suppression (Butterworth et al., 2022; Mele, 2000). The student does not experience a private acknowledgment of wrongdoing that they then conceal; they experience a self-understanding in which the wrongdoing has not yet been established as such. What concealment hides, artificial integrity has not yet formed. This parallels Tenbrunsel and Messick’s (2004) account of ethical fading, in which moral dimensions become progressively less salient through reframing. In artificial integrity, however, the fading is not a gradual erosion but a structurally maintained condition, actively sustained by the design of the systems through which work is produced.

This structural difference generates a testable behavioral signature. A student engaged in concealment should exhibit consistent strategic behavior across contexts, as the private acknowledgment remains stable regardless of the audience (Grigoryeva, 2024). A student in artificial integrity should show context-sensitive variation, with self-understanding destabilizing precisely when forced articulation – a declaration form, a post-submission interview, a prompt history request – makes the contradiction harder to manage (Aquino & Reed, 2002). Disclosure tracks identity threat rather than enforcement severity: highest in anonymous contexts, lowest where explicit specification of cognitive contribution

is required. This pattern – audience-calibrated disclosure, avoidance of specificity, persistent anxiety despite asserted appropriateness – is consistent with managed contradiction.

In confusion-based cases, by contrast, the failure is an informational deficit: the agent lacks the understanding required for accurate informing, and the distinction is not merely taxonomic but diagnostic: it determines what kind of intervention could plausibly restore the informing relationship. The framework does not require that students experience themselves as dishonest; rather, it holds that contradictory knowledge becomes salient under conditions of threat, evaluation, or forced articulation. Behavioral indicators (strategic disclosure, avoidance of process transparency, detection anxiety, defensive justification) are consistent with managed contradiction (Bandura, 1999; Espartinez, 2026a).

The diagnostic implication is direct: if misconduct is neither concealment nor confusion, interventions based on detection or clarification do not merely underperform. They target the wrong mechanism, attempting to compel accurate informing from an agent whose capacity for accurate self-informing has been structurally compromised.

HOW GENAI STABILIZES KNOWING SELF-DECEPTION AS INFORMING INFRASTRUCTURE

GenAI did not invent rationalization; students have long reinterpreted their behavior to maintain a sense of integrity when collaboration rules are ambiguous (Amigud & Hosseini, 2024; Bretag, 2016). What distinguishes artificial integrity is that AI systems externalize and automate this process: rationalization is no longer a difficult internal struggle but a seamless, always-available feature built into the technology. This may alter both how students understand their own work and how they present it to instructors.

This argument does not treat AI assistance as inherently epistemically degrading. The conditions theorized here are specific: they arise when AI design embeds rationalization resources that make accurate self-informing psychologically costly to sustain. AI assistance that makes cognitive contribution explicit and traceable (through visible reasoning trails, attributable process steps, or transparent collaboration records) does not activate these mechanisms. The problem is not AI use but the specific affordance configurations that scaffold the managed coexistence of contradictory knowledge. If those configurations are absent, the informing relationship remains intact.

Three design affordances are theorized to underlie this effect.

Anthropomorphic interaction

- *Informing function:* Conversational interfaces enable users to frame AI use as collaboration, supporting narratives of co-authorship rather than substitution (Espartinez, 2026a). This framing shapes self-informing directly: the agent comes to understand themselves as a collaborator rather than a delegator, rendering the AI's role less accessible to conscious self-accounting.
- *Design origin:* Systems are deliberately optimized for conversational fluency and simulated personhood, embedding assumptions about agency and authorship into the interaction design.
- *Legal implications:* The effects of anthropomorphic framing on self-informing have been empirically documented since early deployment. Developers targeting educational contexts can be understood as operating with constructive notice that such interfaces facilitate collaboration narratives that compromise accurate informing. Under foreseeability-based reasoning, the relevant question is not intent but whether these informing effects were reasonably foreseeable.

Process invisibility

- *Informing function:* Architectural opacity claims that the AI contribution cannot be precisely specified, sustaining ambiguity about authorship. The agent can truthfully claim – even to themselves – that they cannot determine what AI contributed, converting an informing failure into an epistemological excuse.
- *Design origin:* Generative systems are architecturally opaque, providing outputs without transparent accounts of intermediate reasoning (Zerilli et al., 2019). This opacity is not incidental but structural to current architectures.
- *Legal implication:* The same opacity that supports rationalization produces institutional evidentiary limits. Detection systems yield probabilistic outputs, complicating disciplinary processes that require adequate evidentiary grounds. The affordance, therefore, compromises informing integrity at two levels simultaneously: the cognitive (shaping self-informing) and the institutional (shaping the client’s capacity to verify it).

Originality illusions

- *Informing function:* Fluent, novel outputs support the claim that work is “not plagiarism,” even where cognitive labor has been substantially delegated (Espartinez, 2026a). This illusion operates directly on the agent’s self-informing: the output’s apparent originality makes it difficult for the agent to recognize that their own intellectual contribution was minimal, as the work’s surface features resemble those of original authorship.
- *Design origin:* Models are explicitly optimized for coherence, fluency, and apparent originality, with these features treated as quality metrics.
- *Legal implication:* Academic integrity frameworks are largely structured around textual misappropriation. AI-generated content does not map onto this category, creating a gap between recognized plagiarism doctrine and AI-mediated cognitive delegation. From an informing perspective, the client lacks the conceptual vocabulary to name the failure that has occurred – a predictable consequence of optimizing outputs to evade existing detection paradigms.

These three affordances operate as a mutually reinforcing system. Anthropomorphic interaction supplies a collaborative narrative. Process invisibility shields that narrative from specification. Originality illusions provide a rule-compliant surface justification. Jointly, they relocate rationalization from an internal cognitive effort – susceptible to self-correction – to an external, continuously available resource that is theorized to systematically degrade the agent’s capacity for accurate self-informing. The informing relationship is not merely strained; it is structurally inverted.

THE SELF-REINFORCING DYNAMIC: WHEN INFORMING FAILURE BECOMES ENTRENCHED

Artificial integrity is not static. The mechanism is motivated reasoning under reduced resistance: each instance in which AI-assisted work is framed as acceptable, submitted, and validated functions as evidence supporting the rationalization. These episodes accumulate into an evidentiary history – the interpretation has been tested, accepted, and not challenged – which becomes part of the epistemic structure sustaining the self-deception (Tenbrunsel & Messick, 2004). Each client acceptance reinforces the narrative and progressively makes it harder for the agent to access the contradictory knowledge that would require revising it.

Identity processes amplify this effect. As AI-assisted work is repeatedly incorporated into successful academic performance, it becomes integrated into the student’s self-concept as an author (Espartinez, 2026a). Challenges to the behavior are then experienced not as corrections to informing but as threats to identity, prompting defensive justification rather than revision. The agent’s capacity

to inform accurately about their cognitive labor has not merely been compromised; it has been functionally reorganized.

Figure 1 presents artificial integrity as a self-reinforcing system, illustrating how epistemic layering interacts with AI design affordances and institutional conditions to degrade the informing relationship over time.

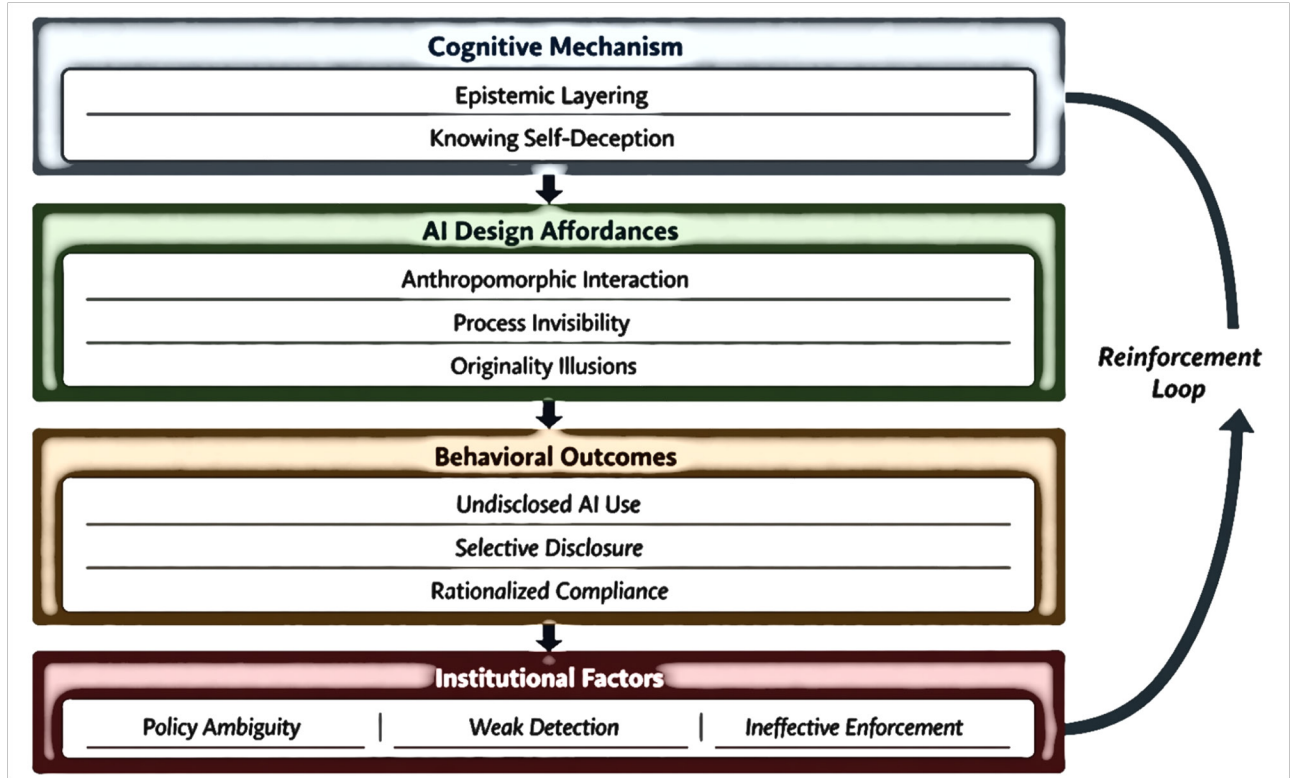


Figure 1. Artificial integrity as a self-reinforcing system

FRAMEWORK IN ACTION: RESOLVING THREE EMPIRICAL PUZZLES AS INFORMING FAILURES

Three behavioral patterns associated with AI-assisted academic work have resisted unified explanation: students who endorse integrity values yet engage in undisclosed AI use; students who vary disclosure by audience; and students who avoid mandatory declaration despite understanding the requirement. Existing integrity measures underperform not because they are poorly implemented, but because they were designed for a failure mode – concealment or confusion – that artificial integrity is not. Concealment and confusion treat these as separate problems, whereas artificial integrity sees a single mechanism. Artificial integrity accounts for them as different manifestations of a single informing failure: the managed coexistence of contradictory knowledge that renders the agent an unreliable informant to both self and client.

THE INTEGRITY-BEHAVIOR PARADOX: WHEN THE AGENT'S INFORMING TO SELF BECOMES UNRELIABLE

Students who strongly endorse academic integrity values nonetheless engage in extensive undisclosed use of AI (Bozkurt et al., 2020; Kirsanov et al., 2025). From an informing perspective, this pattern reveals a fundamental disconnect between two levels of informing: the agent's professed commitments to the client (expressed through value endorsement) and the agent's actual informing about

cognitive labor (expressed through disclosure behavior). Traditional models predict that moral commitment constrains behavior and supports accurate informing; here, it does not. Confusion is also insufficient: students demonstrate accurate knowledge of policies and can identify violations in others' work, indicating that they possess the information necessary for accurate informing.

Artificial integrity accounts for this pattern through motivated moral reinterpretation (the mechanism detailed in the preceding section), by which students reclassify AI-assisted activity into acceptable categories such as "collaboration," "assistance," or "curation," so that self-informing filters out contradictory information rather than eliminating it. The integrity-behavior link is not broken; it is redirected through self-informing. The agent genuinely believes they are acting in accordance with their values because their self-informing has been structured to exclude the contradiction.

That contradictory knowledge remains accessible is revealed by strategic disclosure – concealing AI use in high-stakes contexts while revealing it in low-risk ones – and by detection anxiety despite asserted appropriateness. These responses are inconsistent with genuine belief but consistent with managed contradiction.

THE DISCLOSURE VARIANCE PARADOX: WHEN INFORMING TRACKS IDENTITY THREAT RATHER THAN ACCURACY

Students systematically vary their disclosure of AI use depending on the audience (Espartinez, 2025; Tang, 2025). The information about what actually occurred does not change; what changes is the perceived cost of transmitting it.

If AI use were genuinely appropriate in the student's self-understanding, concealment would be unnecessary. If it were clearly inappropriate and privately acknowledged, one would expect either consistent concealment or consistent disclosure with acknowledgment – neither is observed. Disclosure decisions track perceived threats to identity: disclosure is more likely when it can be incorporated into a validated narrative ("acceptable collaboration") and less likely when it risks exposing the limits of that narrative.

The agent's capacity to vary disclosure strategically confirms that the relevant information is accessible when they choose to provide it. The failure is not one of availability – they know what they did – but of willingness to transmit it when doing so would destabilize self-understanding.

THE DECLARATION AVOIDANCE PARADOX: WHEN FORCED SPECIFICITY DESTABILIZES SELF-INFORMING

Students frequently avoid mandatory AI-use declarations despite understanding the requirement and often asserting that their use was appropriate (Bao & Zeng, 2026). This is not confusion or oversight; they comply with other administrative requirements without difficulty.

Artificial integrity accounts for this avoidance as a response to forced specificity. Informal rationalizations depend on ambiguity. Phrases such as "assistance" or "collaboration" allow flexible interpretation. Formal declarations require explicit articulation (e.g., who performed which cognitive tasks) that exposes the inadequacy of those rationalizations. Students could provide detailed accounts of AI involvement; they avoid doing so because such accounts would collapse the interpretive framework sustaining their sense of integrity. When the client demands specificity – requiring the agent to move from "collaboration" to "I generated the thesis with AI and then rewrote it" – the ambiguity that sustained self-informing dissolves.

UNIFIED EXPLANATION AND ALTERNATIVE ACCOUNTS: INFORMING FAILURE AS A COMMON MECHANISM

Alternative explanations are possible: situational uncertainty may contribute to disclosure variability; social desirability concerns may motivate impression management; context-dependent belief activation may produce shifting justifications without requiring simultaneous contradictory awareness.

Each captures part of the phenomenon. None accounts for all three patterns simultaneously – particularly the combination of strategic variation, specificity avoidance, and sensitivity to potential challenge. Artificial integrity offers a more unified account by locating the mechanism in the management of contradiction rather than its absence or concealment. This does not conclusively establish knowing self-deception as the operative mechanism; it identifies it as the most coherent explanatory framework given the available evidence. The propositions below specify conditions under which the account can be tested and potentially falsified.

TESTABLE PROPOSITIONS: DISTINGUISHING ARTIFICIAL INTEGRITY FROM CONFUSION AND CONCEALMENT

The propositions advanced here are theoretical and diagnostic. Their truth value is an empirical question; the conceptual analysis establishes their plausibility and specifies the conditions under which they could be tested or refuted. The companion article develops five propositions concerning the sociological architecture of artificial integrity (Espartinez, 2026a); the following five address its internal cognitive mechanism, organized around a central distinction: confusion-based failures occur when the agent lacks information (remedy: provision); concealment-based failures occur when the agent withholds information (remedy: deterrence); artificial integrity involves a third type in which the agent possesses information but the organization of self-understanding compromises self-informing (remedy: restructuring the conditions under which self-informing occurs). Table 1 defines the four constructs whose boundaries the framework must maintain – distinguished by cognitive structure, relationship to contradictory knowledge, and behavioral signature – before the propositions are stated.

Table 1. Construct definitions and boundaries

Construct	Cognitive structure	Relationship to contradictory knowledge	Behavioral signature
Concealment	Agent privately acknowledges wrongdoing	Knowledge is present and stable; withheld deliberately	Consistent strategic behavior across contexts, regardless of audience
Rationalization	Agent constructs a post hoc justification for recognized wrongdoing	Knowledge precedes and motivates the justification	Justification deployed after the act; varies with social pressure
Identity protection	Agent manages self-concept under threat	Knowledge is accessible but experienced as destabilizing	Defensive escalation when challenged; narrative elaboration under scrutiny
Knowing self-deception (artificial integrity)	Agent maintains contradictory knowledge without resolving it	Knowledge is retained and accessible but overridden by motivated reinterpretation	Context-sensitive variation; destabilization under forced articulation; anxiety despite asserted appropriateness

Note: These constructs are theoretically adjacent and may co-occur empirically. The boundaries drawn here reflect structural differences in cognitive mechanisms rather than mutually exclusive behavioral categories.

The following propositions represent theoretically derived expectations rather than empirically confirmed relationships and are offered as a basis for future testing. Each states the conditions under which the artificial integrity account would be disconfirmed and an alternative explanation favored.

P1: MOTIVATED AMBIGUITY EXPLOITATION

Artificial integrity predicts that policy ambiguity facilitates undisclosed AI use by providing rationalization space. Students under moderately ambiguous policies should exhibit higher rates of undisclosed AI use than those under highly explicit or absent policies – and, critically, should demonstrate accurate policy understanding when evaluating others’ behavior while selectively invoking ambiguity in self-justification.

P2: DISCLOSURE TRACKS IDENTITY THREAT, NOT DETECTION RISK

If artificial integrity operates through identity management rather than deterrence, disclosure should vary with perceived threat to self-concept rather than enforcement severity: highest in anonymous contexts, moderate with non-evaluative audiences, lower with evaluative instructors, lowest in competitive peer contexts. Disclosure that tracks enforcement intensity rather than audience identity risk would support deterrence-based accounts; disclosure that tracks identity threat independently of enforcement severity would support the artificial integrity account.

P3: PROCESS TRANSPARENCY OUTPERFORMS DETECTION

Transparency interventions – draft histories, reasoning logs, contribution statements – should reduce undisclosed AI use more effectively and durably than detection technologies or mandatory declarations, because they disrupt rationalization by requiring specification of cognitive contribution, making euphemistic categories less sustainable. Detection-based interventions are expected to produce short-term behavioral adjustments in response to enforcement presence, without altering underlying rationalization. If detection consistently outperforms transparency in sustained compliance, this would challenge the artificial integrity account and suggest that the failure is primarily due to concealment.

P4: METACOGNITIVE AWARENESS OUTPERFORMS POLICY INFORMATION

If artificial integrity reflects rationalization deployment rather than informational deficit, training that helps students recognize how ambiguity, euphemistic relabeling, and identity threat shape their reasoning should outperform rule-based instruction. A confusion-based account predicts the opposite: that policy clarification produces equivalent or superior compliance. If metacognitive training produces superior outcomes, this supports the claim that the failure lies not in what the agent knows but in how self-informing is organized to manage what is known.

P5: LATE-STAGE INTERVENTION REQUIRES FACE-SAVING PATHWAYS

Artificial integrity predicts that repeated validation of rationalizations produces identity commitment effects that increase resistance to direct confrontation over time. Students with extensive histories of undisclosed AI use should respond to detection or accusation with intensified justification rather than behavioral change. Interventions offering face-saving pathways – allowing reinterpretation of past behavior without requiring explicit admission of wrongdoing – should reduce resistance by lowering the identity cost of revising self-understanding. If direct confrontation consistently produces compliance without defensive escalation, this would falsify the artificial integrity account and support a concealment-based explanation.

LEGAL, REGULATORY, AND DESIGN IMPLICATIONS

Artificial integrity cannot be treated solely as a problem of student conduct. Its causes and consequences are distributed across developers, institutions, and regulatory frameworks – each of which participates in the informing infrastructure that shapes whether agents can inform accurately about their cognitive labor. The breakdown in the client–agent relationship between students and institutions is not a failure of individual compliance but a failure of informing infrastructure, and restoring it requires redesigning that infrastructure at every level.

DEVELOPER RESPONSIBILITY: FROM FORESEEABLE HARM TO DESIGN ETHICS

The affordances that scaffold artificial integrity – anthropomorphic interaction, process invisibility, and originality illusions – are the result of deliberate design choices: conversational interfaces optimized for simulated personhood, architecturally opaque inference processes, and training objectives that prioritize fluent, novel output.

The question is not whether these features were intended to facilitate misconduct, but whether their downstream effects on informing integrity were reasonably foreseeable. Product liability reasoning provides the relevant normative framework. Under the Restatement (Third) of Torts: Products Liability, designers may be held responsible for harms arising from foreseeable uses of their products even if unintended. The argument here is not that existing doctrine straightforwardly applies to GenAI; rather, that its underlying logic, foreseeability as a basis for responsibility, offers a conceptually appropriate foundation for extension to informing infrastructure.

Constructive notice was established no later than 2023, when the first empirical wave documented how conversational interfaces scaffold collaboration narratives and how AI-generated outputs complicate attribution of intellectual labor. Developers deploying systems in educational contexts after that date operate with knowledge, or at a minimum constructive notice, that these features enable rationalization architectures that systematically degrade the integrity of information. Three questions follow that the law-ethics-technology literature can no longer defer: whether existing product liability frameworks extend to AI systems whose features foreseeably scaffold psychological conditions that compromise informing integrity; what design modifications discharge the duty of care – transparency tools, de-anthropomorphized interfaces, output provenance markers, pre-deployment integrity impact assessments; and whether educational AI systems should require informing integrity impact assessments as a condition of deployment, analogous to GDPR-mandated data protection impact assessments (Intersoft Consulting, n.d.). This analysis aligns with the EU AI Act (Smuha, 2025), which classifies AI systems used in educational and vocational training contexts as high-risk under Annex III, triggering mandatory conformity assessments, transparency obligations, and human oversight requirements. The rationalization affordances identified in this paper are precisely the design-level informing risks that a conformity assessment framework ought to capture. Developers who deploy in educational markets without assessing them are not merely negligent; they are building infrastructure that systematically undermines the conditions for accurate informing.

REGULATORY INADEQUACY: THE BINARY THAT ARTIFICIAL INTEGRITY DISSOLVES

Current academic integrity frameworks are structured around a binary that artificial integrity dissolves. Regulatory routines address either deliberate concealment – agents who know they have violated rules and hide it – or genuine confusion – agents who misunderstand what the rules require. Artificial integrity describes neither. Students engaging in it accurately understand institutional rules, do not privately experience themselves as violating them, yet retain accessible knowledge that their justificatory narratives would not survive scrutiny. No existing regulatory category captures this informing failure. That absence is not a technical gap waiting to be filled; it is a structural failure of regulatory imagination.

The legal consequences are concrete. Detection-based enforcement presupposes concealment and misidentifies the target: it treats the informing failure as willful misinforming when the agent does not experience themselves as having information to hide. Policy clarification presupposes confusion and supplies vocabulary that students incorporate into rationalizations rather than constraints on behavior. Punitive sanctions presuppose deterrable deliberate wrongdoing and escalate defensive justification, precisely the identity-protection response the framework predicts.

Goss v. Lopez (Justia U.S. Supreme Court, 1975) requires that sanctions be supported by adequate evidence. Where the only available evidence is probabilistic detection output, the evidentiary basis is not merely weak but structurally absent: a tool that identifies AI-generated text cannot identify AI-delegated cognitive labor that has been revised out of the final submission. The student who generates a thesis with AI, rewrites it in their own prose, and submits the result has committed no textual act any detector can find.

Dixon v. Alabama (1961) exposes a definitional problem that precedes any procedural one. Artificial integrity cannot currently be charged with the specificity that adequate notice demands. A charge of submitting AI-generated text is one the student can truthfully deny – the submission is not AI-generated in any detectable sense. A charge of outsourcing cognitive labor confronts a different obstacle: no legal framework yet defines what constitutes adequate cognitive contribution, what evidence would demonstrate it, or where permissible AI assistance ends and prohibited cognitive delegation begins. Existing disciplinary language cannot yet name the informing failure with sufficient precision for a meaningful response.

Schaer v. Brandeis University (2000) adds a contractual dimension. Ambiguity in student codes does not merely permit rationalization; under Schaer's contract-interpretation framework, it arguably creates enforceable interpretive latitude for the student who deploys it. AI-use provisions drafted in terms like "unauthorized assistance," "responsible use," or work that must "reflect the student's own thinking" are precisely the language artificial integrity exploits. A student arguing that their understanding of "collaborative AI use" falls within those terms, as reasonably interpreted, would have a colorable claim – particularly where the institution has tacitly tolerated AI assistance in practice. Institutional ambiguity is therefore not only a normative failure; under Schaer, it is a contractual vulnerability that undermines the institution's capacity to demand accurate informing.

A legally adequate framework would need to accomplish three things current frameworks do not: recognize knowing self-deception as a distinct third category of informing failure requiring different evidentiary standards and remedial approaches; assign responsibility not only to individual agents but to the institutional and technological conditions that make informing failure sustainable; and design remedies that collapse rationalization space through process transparency, contribution documentation, and metacognitive intervention, rather than intensifying enforcement pressure on a condition they misdiagnose.

INSTITUTIONAL RESPONSIBILITY AND THE DUTY OF CARE

The conditions under which artificial integrity stabilizes are expected to vary by institutional culture, assessment design, and disciplinary context – and this variation is itself theoretically informative. Where the process is visible, expectations are clear, and AI policies are consistently enforced, the rationalization space the framework describes contracts; alternative accounts, including genuine confusion or sincere belief revision, become more plausible.

Institutions are not neutral actors. Policies that preserve ambiguity through terms like "appropriate assistance," "collaboration," or work that must reflect "the student's own thinking" – function as resources for rationalization rather than constraints on behavior. Where such ambiguity persists despite evidence of its effects, institutions are not merely failing to prevent informing failure; they are actively supplying the conditions that sustain it. Ambiguity that predictably enables informing failure, preserved by institutions with the capacity to reduce it, may reasonably be characterized as institutional complicity in the failure they purport to address. Whether this constitutes a legally cognizable duty of care – to students whose learning is undermined, to employers and credential-relying parties who depend on accurate accounts of cognitive labor, or to the integrity of credentialing systems more broadly – is a question the law-ethics-technology tradition is well placed to pursue.

Three institutional design principles follow directly. First, consistency of enforcement matters more than severity: inconsistent application validates excuse construction and reinforces the plausibility of

self-justifying interpretations. Second, assessment design is informing infrastructure; tasks that obscure the locus of thinking generate rationalization space; tasks that require visible cognitive contribution constrain it. Third, process-level clarity must take priority over outcome-based evaluation, making cognitive contribution attributable so agents can inform accurately without destabilizing self-understanding. Whether these obligations remain normative recommendations or harden into legal duties will depend on how courts and regulators respond to the evidentiary and definitional gaps this framework exposes – but the direction is clear.

INTERVENTION IMPLICATIONS: COLLAPSING RATIONALIZATION SPACE

Effective intervention requires targeting the conditions that sustain rationalization, not intensifying enforcement against its outcomes. Table 2 maps each AI affordance to its corresponding intervention, each restoring a dimension of the informing relationship that the affordance has compromised: metacognitive training exposes the mechanisms of self-serving reinterpretation; transparency requirements make cognitive labor visible and attributable; process-focused assessment shifts the client’s information demand from “*Is this original?*” (which AI-generated outputs satisfy) to “*How was this made?*” (which requires accurate informing about cognitive labor).

Table 2. Intervention mapping to AI affordances in artificial integrity

AI affordance	Proposed intervention	Informing function restored
Anthropomorphic Interaction (“Collaboration” narratives)	Metacognitive Training (rationalization recognition workshops)	Restores capacity for accurate self-informing by exposing how conversational interfaces scaffold self-serving reframing; interrupts rationalization at the point of construction before it stabilizes self-understanding.
Process Invisibility (“Unmeasurable contribution” claims)	Transparency Requirements (process logs, draft histories, contribution statements)	Makes cognitive labor visible and attributable; removes the ambiguity agents exploit to maintain self-informing while providing incomplete information to the client.
Originality Illusions (“Not plagiarism” defenses)	Process-Focused Assessment (graded drafts, oral defenses, annotated bibliographies)	Shifts evaluation from textual originality – which AI-generated outputs satisfy – to visible cognitive contribution, restoring the client’s capacity to assess actual intellectual labor.

These interventions also address the evidentiary and definitional gaps identified by the legal analysis. Process documentation generates the cognitive-contribution records that disciplinary proceedings require under Goss and Dixon: the client gains reliable information about cognitive labor, and the agent can provide it through a routine process rather than forced admission. Clearer expectation-setting reduces the interpretive ambiguity that, under Schaer, gives students enforceable latitude. Metacognitive training functions as a disclosure intervention – reducing the gap between what students know and what they are willing to articulate, making voluntary transparency more likely before formal proceedings become necessary. Enforcement does not become unnecessary under this model, but its effectiveness depends on prior infrastructure: without mechanisms that make cognitive contributions observable, it will continue to target symptoms. Collapsing rationalization space is simultaneously a pedagogical and a regulatory strategy – one that addresses the structural condition existing frameworks misdiagnose rather than intensifying pressure on it.

CONCLUSION

This paper has argued that artificial integrity is best understood not as concealment or confusion, but as a structural informing failure with cognitive, institutional, and legal dimensions. Students who engage with it retain the knowledge that AI has performed substantial intellectual work, yet maintain a morally acceptable self-understanding as authors, neither deliberately misinforming the client nor lacking the information to inform accurately. Their capacity for accurate self-informing has been compromised by the organization of self-understanding itself. That compromise cascades into unreliable informing to the institutions that depend on truthful accounts of cognitive labor. The mechanism – knowing self-deception sustained through epistemic layering and stabilized by design – offers an account not just of individual behavior but of a systematic failure of the informing infrastructure on which credentialing depends.

Responsibility for that failure cannot be located solely at the level of student conduct. Developers design systems whose affordances foreseeably shape what information is accessible and how it may be integrated into self-understanding; to the extent those effects were foreseeable, foreseeability supplies a principled basis for responsibility. Institutions preserve policy ambiguity that supplies rationalization space rather than constraining it; where that persists despite evidence of its effects, institutions become participants in the failure they purport to regulate. Regulatory frameworks, structured around a concealment-or-confusion binary, are unprepared to address a third type of informing failure – one in which the agent’s capacity for accurate self-informing has been structurally compromised. Recognizing artificial integrity as a distinct category is therefore not a conceptual refinement but a regulatory necessity.

The five propositions advanced here are not an admission of incompleteness; they are an invitation to empirical engagement with a framework that generates testable, falsifiable predictions. Future research should examine how institutional design affects the stability of artificial integrity, whether process-transparency interventions produce sustained improvements over detection-based approaches, how metacognitive training affects the capacity for accurate self-informing, and how courts and regulators respond to the evidentiary and definitional gaps this framework exposes. The legal arguments outline a normative direction rather than a settled doctrinal position; that direction will be refined by how litigation and regulation develop in response to the gaps identified here. Artificial integrity is proposed here as a testable conceptual category within informing science, subject to revision as empirical evidence accumulates.

The central issue is not whether students are honest; it is how honesty remains cognitively and institutionally sustainable under conditions of delegated cognition. When students fail to inform themselves about their own cognitive labor, they also fail to inform the institutions that depend on truthful accounts of that labor. When institutions court ambiguity that enables rationalization, they become complicit in the degradation of the relationship that informs them. In educational contexts where these conditions converge, developers who design systems whose affordances foreseeably scaffold self-informing failure shape, without yet being accountable for, the epistemic conditions under which students learn to account for their own cognitive labor.

So long as institutions preserve ambiguity that predictably enables rationalization, and regulatory frameworks address a binary misconduct model that artificial integrity dissolves, the condition will persist, not as a student failing, but as a structural feature of an ecology that has not yet reckoned with what it has built. Restoring the informing relationship requires designing systems that make cognitive labor visible, crafting policies that collapse rather than supply rationalization space, and developing regulatory frameworks that recognize knowing self-deception as a distinct informing failure requiring distinct remedies. The stakes extend beyond academic integrity: the informing relationship between learner and institution is a model for every professional and civic context in which accurate accounts of cognitive labor are essential. Addressing artificial integrity is therefore not merely about preserving academic standards. In the institutional contexts where it thrives – where AI assistance

shades into cognitive delegation, where policies supply rationalization space rather than constraining it, and where outputs are evaluated without regard for the process that produced them – it is about preserving the conditions under which accurate self-informing remains possible at all.

REFERENCES

- Alsharefeen, R., & Al Sayari, N. (2025). Examining academic integrity policy and practice in the era of AI: A case study of faculty perspectives. *Frontiers in Education*, 10, 1621743. <https://doi.org/10.3389/feduc.2025.1621743>
- Amigud, A., & Hosseini, S. (2024). Collaboration, collusion, and barter-cheating: An analysis of academic help-seeking behaviors. *Assessment & Evaluation in Higher Education*, 49(3), 392–407. <https://doi.org/10.1080/02602938.2023.2259631>
- Aquino, K., & Reed, A. (2002). The self-importance of moral identity. *Journal of Personality and Social Psychology*, 83(6), 1423–1440. <https://doi.org/10.1037/0022-3514.83.6.1423>
- Bandura, A. (1999). Moral disengagement in the perpetration of inhumanities. *Personality and Social Psychology Review*, 3(3), 193–209. https://doi.org/10.1207/s15327957pspr0303_3
- Bandura, A. (2018). A commentary on moral disengagement: The rhetoric and the reality. *The American Journal of Psychology*, 131(2), 246–251. <https://doi.org/10.5406/ameripsyc.131.2.0246>
- Bao, A., & Zeng, Y. (2026). AI disclosure, moral shame, and the punishment of honesty. *Accountability in Research*, 33(3), 2542197. <https://doi.org/10.1080/08989621.2025.2542197>
- Bobier, C., Rodger, D., Gillham, A., Dutta, P., & Merlocco, A. (2025). The missing disclosure: Is generative AI use in bioethics scholarship going largely unreported? *BMC Medical Ethics*, 26, Article 180. <https://doi.org/10.1186/s12910-025-01349-4>
- Bozkurt, A., Jung, I., Xiao, J., Vladimirsch, V., Schuwer, R., Egorov, G., Lambert, S., Al-Freih, M., Pete, J., Don Olcott, J., Rodes, V., Aranciaga, I., Bali, M., Alvarez, A. J., Roberts, J., Pazurek, A., Raffaghelli, J. E., Panagiotou, N., de Coëtlogon, P., ... Paskevicius, M. (2020). A global outlook to the interruption of education due to COVID-19 pandemic: Navigating in a time of uncertainty and crisis. *Asian Journal of Distance Education*, 15(1), Article 1.
- Bretag, T. (2016). *Handbook of academic integrity*. Springer. <https://doi.org/10.1007/978-981-287-098-8>
- Butterworth, J., Trivers, R., & von Hippel, W. (2022). The better to fool you with: Deception and self-deception. *Current Opinion in Psychology*, 47, 101385. <https://doi.org/10.1016/j.copsyc.2022.101385>
- Cohen, E. (1999). Reconceptualizing information systems as a field of the transdiscipline informing science: From ugly duckling to swan. *Journal of Computing and Information Technology*, 7(3), 213–219.
- Cohen, E. (2009). A philosophy of informing science. *Informing Science: The International Journal of an Emerging Transdiscipline*, 12, 1–15. <https://doi.org/10.28945/425>
- Cotton, D. R. E., Cotton, P. A., & Shipway, J. R. (2024). Chatting and cheating: Ensuring academic integrity in the era of ChatGPT. *Innovations in Education and Teaching International*, 61(2), 228–239. <https://doi.org/10.1080/14703297.2023.2190148>
- Dawson, P. (2021). *Defending assessment security in a digital world*. Routledge. <https://doi.org/10.4324/9780429324178>
- Dixon v. Alabama, 294 F.2d 150, 5th Cir. (1961). <https://law.justia.com/cases/federal/appellate-courts/F2/294/150/346079/>
- Espartinez, A. (2025). Bridging the educational divide with ChatGPT's integration in Philippine higher education: Q methodology and narrative inquiry studies. *Journal of Information Technology Education: Innovations in Practice*, 24, 11. <https://doi.org/10.28945/5526>
- Espartinez, A. (2026a). Artificial integrity: Social performativity of honesty in the age of generative AI. *Law, Ethics & Technology*, 3(1), 1–19. <https://doi.org/10.55092/let20260003>

- Espartinez, A. (2026b). Informing for equity: A systemic framework for GenAI assessment in resource-constrained higher education. *Informing Science: The International Journal of an Emerging Transdiscipline*, 29, Article 4. <https://doi.org/10.28945/5734>
- Festinger, L. (1957). *A theory of cognitive dissonance*. Stanford University Press. <https://doi.org/10.1515/9781503620766>
- Gill, G., Mullarkey, M., & Satterfield, R. K. (2018). Informing on a rugged landscape: How complexity drives our preferred information sources. *Informing Science: The International Journal of an Emerging Transdiscipline*, 21, 1-18. <https://doi.org/10.28945/3977>
- Goffman, E. (1959). *The presentation of self in everyday life*. Doubleday.
- Gonsalves, C. (2025). Addressing student non-compliance in AI use declarations: Implications for academic integrity and assessment in higher education. *Assessment & Evaluation in Higher Education*, 50(4), 592–606. <https://doi.org/10.1080/02602938.2024.2415654>
- Grigoryeva, M. S. (2024). A theory of concealment. *Theory and Society*, 53, 1321–1355. <https://doi.org/10.1007/s11186-024-09578-z>
- Intersoft Consulting. (n.d.). *Art. 35 GDPR: Data protection impact assessment*. <https://gdpr-info.eu/art-35-gdpr/>
- Justia U.S. Supreme Court. (1975). *Goss v. Lopez*, 419 U.S. 565 (1975). <https://supreme.justia.com/cases/federal/us/419/565/>
- Kirsanov, O., Kushwah, L., & Selvaretnam, G. (2025). Beyond detection: How students use – and hide – AI in online assessments and what authentic tasks can do about it. *Journal of Academic Ethics*, 24, Article 14. <https://doi.org/10.1007/s10805-025-09691-3>
- Kunda, Z. (1990). The case for motivated reasoning. *Psychological Bulletin*, 108(3), 480–498. <https://doi.org/10.1037/0033-2909.108.3.480>
- Leary, M. R., & Kowalski, R. M. (1990). Impression management: A literature review and two-component model. *Psychological Bulletin*, 107(1), 34–47. <https://doi.org/10.1037/0033-2909.107.1.34>
- Mele, A. (2000). *Self-deception unmasked*. Princeton University Press. <https://doi.org/10.1515/9781400823970>
- Nahl, D., & Bilal, D. (2007). *Information and emotion: The emergent affective paradigm in information behavior research and theory*. Information Today, Inc.
- Resnik, D. B., & Hosseini, M. (2026). Disclosing artificial intelligence use in scientific research and publication: When should disclosure be mandatory, optional, or unnecessary? *Accountability in Research*, 33(2), Article 2481949. <https://doi.org/10.1080/08989621.2025.2481949>
- Rudolph, J., Tan, S., & Tan, S. (2023). ChatGPT: Bullshit spewer or the end of traditional assessments in higher education? *Journal of Applied Learning and Teaching*, 6, 342–363. <https://doi.org/10.37074/jalt.2023.6.1.9>
- Schaer v. Brandeis University, 432 Mass. 474. (2000). https://bio-tech.law.lsu.edu/cases/schools/schaer_v_brandeis.htm
- Smuha, N. A. (2025). Regulation 2024/1689 of the European Parliament & Council of June 13, 2024 (EU Artificial Intelligence Act). *International Legal Materials*, 64(5), 1234–1381. <https://doi.org/10.1017/ilm.2024.46>
- Tang, B. L. (2025). Undeclared AI-assisted academic writing as a form of research misconduct. *Science Editor*. <https://doi.org/10.36591/SE-4804-02>
- Tenbrunsel, A. E., & Messick, D. M. (2004). Ethical fading: The role of self-deception in unethical behavior. *Social Justice Research*, 17(2), 223–236. <https://doi.org/10.1023/B:SORE.0000027411.35832.53>
- Tsang, J.-A. (2002). Moral rationalization and the integration of situational factors and psychological processes in immoral behavior. *Review of General Psychology*, 6(1), 25–50. <https://doi.org/10.1037/1089-2680.6.1.25>
- Weber-Wulff, D., Anohina-Naumeca, A., Bjelobaba, S., Foltýnek, T., Guerrero-Dib, J., Popoola, O., Šigut, P., & Waddington, L. (2023). Testing of detection tools for AI-generated text. *International Journal for Educational Integrity*, 19, Article 26. <https://doi.org/10.1007/s40979-023-00146-z>

Zerilli, J., Knott, A., Maclaurin, J., & Gavaghan, C. (2019). Algorithmic decision-making and the control problem. *Minds and Machines*, 29(4), 555–578. <https://doi.org/10.1007/s11023-019-09513-7>

AUTHOR



Dr. Espartinez is an Assistant Instructional Professor in Philosophy at Providence College in Rhode Island, USA, and an adjunct at De La Salle–College of Saint Benilde in Manila, Philippines. With a background in Philosophy, including a PhD from the University of Santo Tomas, Dr. Espartinez is a recognized authority in the field. Her expertise extends to ethics, the philosophy of the human person, artificial integrity, and pedagogy. She has authored philosophy textbooks, published articles, and delivered notable papers. Her academic contributions earned her prestigious awards, including the Fulbright in Residence and Visiting Professorship, as well as government grants for innovation.