# ASCILITE 2026

***The Future is Connection:***

*Collaboration, Creativity and Compassion in a Digital Society*


# Caught in the algorithm: Student discourse on AI governance, surveillance, and academic integrity in higher education


**Preya Shah, Harsh Patel, Suren Ladd**

James Cook University



This study examines how awareness of generative AI academic integrity policies shapes student anxiety, ethical reasoning, and writing practices in higher education. Drawing on a reflexive thematic analysis of student discourse from YouTube and Reddit, the study explores how detection-oriented governance frameworks are experienced and negotiated outside formal institutional contexts. Findings indicate that students do not primarily describe intentional rule-breaking; instead, they articulate experiences of surveillance, uncertainty, and adaptation within assessment systems perceived as inconsistent and adversarial. Three interrelated themes are identified: surveillance anxiety and defensive authorship; normative dissonance and moral negotiation of AI use; and cognitive offloading, systemic strain, and declining trust in assessment integrity. Across these themes, AI policy awareness is shown to shape not only behavioural responses but also students' affective and epistemic relationships with writing, authorship, and institutional authority. The study argues that detection-first approaches risk intensifying anxiety and undermining trust without resolving underlying tensions in assessment design. It concludes that generative AI exposes structural fragilities in higher education, suggesting the need to reorient academic integrity governance toward design-led, transparent, and equity-sensitive approaches that prioritise learning, trust, and student experience.



*Keywords:* academic integrity, GenAI, student wellbeing, surveillance anxiety, cognitive offloading


## Introduction

The emergence of generative AI has prompted considerable debate about academic integrity, skill development, and assessment in higher education (McGuirk, 2025). Much of this discussion has focused on detection, deterrence, and policy enforcement, often obscuring how students experience and respond to these systems. Research consistently demonstrates that ethical decision-making is socially situated and cannot be reduced to rule compliance alone. Where institutional responses are experienced as punitive or surveillance-oriented, students may become more concerned with managing risk than engaging with ethical principles. Generative AI can also facilitate cognitive offloading, substituting for effortful cognitive processes involved in reasoning, writing, and problem-solving (Gerlich, 2025; Risko & Gilbert, 2016), raising concerns about disciplinary thinking and students' sense of authorship. AI detection technologies have further generated equity concerns, with evidence suggesting these systems produce false positives and disproportionately affect multilingual writers and neurodivergent students (Cotton et al., 2024, Perkins et al., 2023), extending their implications beyond technical accuracy to questions of trust and inclusion. Student perspectives on these conditions, however, remain underexplored in the existing literature. This study analyses student discourse on social media to examine how AI governance is experienced and contested beyond institutional settings, conceptualising detection-oriented governance as institutional surveillance that shapes student identity, writing practices, and trust in assessment, positioning responses to AI as negotiations of identity, fairness, and agency in algorithmically mediated education.

## Literature review

Current academic integrity scholarship has increasingly turned its attention to the conditions under which students make decisions about AI use, moving beyond compliance-focused models toward contextual and relational accounts. Selwyn et al. (2020) argue that students construct coherent ethical frameworks around AI use that institutions frequently fail to recognise, framing non-compliance less as moral failure than as rational adaptation to structural conditions. Similarly, Moya Figueroa et al. (2024) find that punitive integrity frameworks position students as suspects, encouraging avoidance over learning and deepening rather than

resolving trust deficits. Cotton et al. (2024) demonstrate that AI use decisions are shaped as much by workload, ambiguity, and perceived inequity as by ethical reasoning. Recent research further highlights the affective dimensions of this governance landscape: Bearman et al. (2025) found that students experienced marked anxiety about AI use when uncertain whether it constituted cheating; while algorithmic surveillance has been theorised as an affective engine of control that compels students toward authorial alienation and reinforces linguistic injustice, particularly for non-native English writers (supported by Liang et al., 2023). At the cognitive level, Risko and Gilbert (2016) theorise cognitive offloading as the displacement of effortful cognitive processes onto external tools, a dynamic Gerlich (2025) identifies as increasingly relevant to AI-assisted writing. Boud and Soler (2016) and Bearman and Ajjawi (2019) argue that process-oriented and authentic assessment approaches better support sustained learning than output-focused evaluation, a finding with direct implications for AI governance design.

## Methods

The study investigates how students experience and negotiate AI academic integrity policies outside formal institutional settings, focusing on naturally occurring public discourse from Reddit and YouTube. These platforms were selected because they host open, voluntary, and unsolicited discussions about generative AI and academic life, providing authentic accounts uninfluenced by researcher presence. A passive observational approach was adopted to minimise the social desirability bias associated with surveys and interviews. Data were collected over six months (November 2025–April 2026) from an initial pool of approximately 40 YouTube videos and 80 Reddit threads identified through keyword searches organised into thematic clusters. YouTube videos were retained based on relevance to AI and academic integrity, student-centred commentary, and a minimum of 20 comments. Reddit threads were selected from r/ChatGPT, r/Advice, and r/Professors based on relevance to AI use in academic assessment, thread depth, and identifiable student perspectives. Sampling was guided by four keyword clusters: AI detection and misconduct, AI tools and productivity, writing and cognition, and institutional policy. Keywords such as Turnitin, ChatGPT, critical thinking, and academic integrity were used to retrieve relevant data. The final corpus comprised 1,241 YouTube comments and 2,743 Reddit comments, totalling 3,984 units of discourse ($n$ = 3,984).

Data were analysed using inductive thematic analysis. Analysis began with repeated reading to achieve familiarity with the data, after which initial codes were generated by identifying recurring ideas and patterns relating to AI governance and academic integrity. Codes were refined into broader themes through constant comparison across both platforms and reviewed for internal coherence, conceptual distinction, and alignment with the research question. Representative quotations were selected to illustrate each theme while preserving participants' language and context. All data were publicly available and non-identifiable; usernames were omitted and quotations reported only where individuals could not reasonably be identified, consistent with ethical guidance for passive analysis of public online content (Moreno et al., 2013, p. 709). As the study involved passive analysis of publicly available, non-identifiable data with no researcher interaction with participants, it met the criteria for exemption from formal ethics committee review under institutional guidelines for low-risk internet research. As researchers working within higher education, we hold views on AI governance and academic integrity that may have shaped our interpretive sensibilities during coding. To manage this, we engaged in ongoing reflexive discussion throughout the analytical process, documenting coding decisions and revisiting initial interpretations to distinguish student-reported experience from researcher inference. Disagreements were resolved through return to the data rather than consensus by default.

## Findings

The analysis yielded three interrelated themes collectively illustrating how students experience, rationalise, and are affected by AI detection and policy within higher education.

### Theme 1: Surveillance anxiety defensive, authorship, and equity

The findings indicate that students' primary emotional response to AI detection is not guilt but anxiety about being wrongly accused by systems they perceive as arbitrary, opaque, and unreliable. Accounts do not suggest

dishonest intent; rather, they highlight widespread concerns about systems perceived as arbitrary, opaque, and unreliable. One student explained, "I don't want to be accused of plagiarism based off of vibes and I thought academics would have a bit more rigor than just trusting a program that hasn't been proven to work" (R1, Reddit); while another admitted being too afraid to submit work without first using an AI humaniser tool (R2, Reddit). In more acute cases, uncertainty became psychologically overwhelming, with one student receiving a high detection score for entirely original work and writing simply: "I can't do this anymore" (R3, Reddit). These accounts suggest not recklessness but hypervigilance, with students investing considerable emotional energy managing a risk beyond their control, aligning with Moya Figueroa et al. (2024), who found that detection-focused cultures shift attention from learning towards risk management and compliance. These anxieties were not confined to academic forums. YouTube viewers expressed similar frustrations: "I have written the content by myself...but when I checked in ZeroGPT...it is showing AI generated...couldn't understand why?" (YT1, YouTube); "Recently I wrote a poem, when I scanned it says 81% by AI, even I didn't use AI to write my poem" (YT2, YouTube); and one commenter reported a paper completed before ChatGPT existed being flagged as 82% AI-generated, adding: "I just put effort into grammar etc" (YT3, YouTube). Detection anxiety also influenced composing practices in ways that may undermine authentic academic development. Students described modifying their writing to appear less AI-like by simplifying vocabulary, shortening sentences, or introducing grammatical inconsistencies: "I am a good writer...but now I have to do mistakes on purpose" (R4, Reddit). Equity concerns further shaped these experiences, as neurodivergent students and those with distinctive writing styles reported more frequent flagging linked to features such as em dashes, semicolons, and syntactic precision. Detection frameworks calibrated to normative expression may therefore disadvantage linguistic individuality, raising important questions about fairness in assessment contexts where diverse writing deserves recognition rather than suspicion.

### Theme 2: Normative dissonance and the moral economy of AI use

The findings indicate that where institutions frame AI use as an integrity violation, students largely do not; constructing instead their own moral frameworks that position AI as a pragmatic, and often justified, response to structural pressures. For many, AI was not construed as cheating but as a pragmatic response to exhaustion, competing demands, and perceived institutional unfairness. Arguments about democratisation recurred, with students framing AI as levelling a playing field historically shaped by private tutoring and paid essay services: "Rich students were already commonly paying other people to write their papers. Now that the peasants have access to this, time to fix it!" (R5, Reddit). Others questioned the logic of prohibition, likening it to banning calculators in mathematics (R6, Reddit). These responses point to unresolved tensions in institutional definitions of acceptable academic support that policy frameworks have been slow to address. YouTube comment sections reflected similarly contested moral positions. One viewer argued that enforcement amounted to policing an arbitrary convention: "Whoever spends their time and energy hunting down texts made with ChatGPT is displaying a psychological inability to manage their own battles...Why wouldn't I do it? To preserve some 'sacred' convention?" (YT4, YouTube). Another countered: "If you can't do the work without AI, then you can't think well enough on your own. Comparing it to calculators isn't the same thing" (YT5, YouTube). A third grounded their justification in material necessity: "Anyone who complains about students using this has never had to work three jobs to pay for classes unrelated to their major...I'll never feel bad for this. It's a wonderful tool to cut down on the time it takes for busywork" (YT6, YouTube).

These positions ranging from moral critique of prohibition to defence of cognitive standards, to appeals to structural necessity, mirror the ethical framings evident in the Reddit data. Perceived institutional inconsistency further shaped these interpretations. Students who encountered recycled lectures or limited feedback questioned the legitimacy of strict AI prohibitions, reading such asymmetries as institutional bad faith. These perceptions were reinforced by social normalisation, with the widespread belief that "everyone uses it" rendering non-use both unrealistic and strategically disadvantageous. This aligns with Selwyn et al. (2020, p. 2), who argue that students construct coherent ethical frameworks around AI use that institutions often fail to recognise. The findings point less to a values deficit than to a trust deficit, with direct implications for institutional governance and policy design.

### Theme 3: Cognitive offloading, systemic strain, and the arms race

The findings suggest that sustained AI use is reshaping students' relationship with their own thinking, with

many describing a shift from generative writing to curation; reviewing and submitting AI-produced content rather than developing independent arguments. Writing, once experienced as intellectual struggle and development, was increasingly framed as a supervisory activity. As one student noted, a single prompt left them "pretty much 70% of the way to a complete essay" (R7, Reddit), a convenience perceived as rational under academic pressure yet viewed with concern. One student reflected, "I no longer know how to write essays with my own voice" (R8, Reddit), while another observed that outsourcing thinking weakens reasoning over time (R9, Reddit). These accounts indicate awareness of the cognitive costs of AI reliance, even among students who find such tools difficult to avoid. YouTube commenters expressed similar concerns. One noted that AI "renders the younger generation completely incapable of learning anything, retaining anything, or producing any thoughts of their own" (YT7, YouTube), while another described it as "a cure for lazy minds that in the end will only make you more stupid, as your thinking process atrophies" (YT8, YouTube). A more measured position emphasised conditional use: "It is good to proofread your original work with ChatGPT, but do not stop your brain from thinking. Also, remember that there is a taste of your own tone in research writing" (YT9, YouTube), capturing the tension between augmentation and dependency.

These accounts align with Risko and Gilbert's (2016, p. 676) concept of cognitive offloading, whereby reliance on external tools reduces engagement in the effortful cognitive processes underpinning disciplinary reasoning and epistemic independence. At the systemic level, students situated these developments within an escalating institutional arms race: institutions deploy detection tools, students adapt, and institutions respond with further enforcement. One student observed, "Restricting AI is an arms race that universities and industry will never be able to win" (R10, Reddit). Detection systems were widely described as inconsistent, flagging original work while failing to identify AI-generated submissions. Research confirms this unreliability, with studies combining academic judgement and software finding significant limitations in GPT-4 detection accuracy (Perkins et al., 2023). This was echoed on YouTube: "As new AI tools are coming in market, the AI detectors are also advancing day by day" (YT10, YouTube); "All AI text is initially human written...so technically nothing AI generated is fully AI generated" (YT11, YouTube); "AI detection is a scam. All they know is pattern, nothing to do with an artistic mind" (YT12, YouTube). This perceived unreliability erodes trust in detection technologies, assessment processes, and institutional adjudication. Collectively, these accounts suggest that the most significant consequences of current AI governance may not be deterrence, but declining confidence in assessment systems, reduced engagement in independent writing, and weakened institutional trust.

## Discussion

Across the themes, a consistent pattern emerges: detection-oriented AI governance does not produce ethical alignment. Instead, it generates anxiety, negotiation, and scepticism. Students operate under ambiguous expectations, perceive detection systems as unreliable, and experience misclassification as opaque and uneven. This aligns with Moya Figueroa et al. (2024), who argue that punitive integrity frameworks position students as suspects and encourage avoidance over learning; and with Cotton et al. (2024) who show that AI use decisions are shaped by workload, ambiguity, and perceived inequity as much as by ethical reasoning. Students rarely resist or comply outright; more often, they adapt and negotiate institutional expectations under uncertainty. Detection-centred governance may also displace formative dialogue and authentic development. Students described suppressing stylistic complexity, adjusting writing to anticipated algorithmic patterns, and experiencing assessment as adversarial. Process-oriented approaches such as portfolios, reflective tasks, and staged submissions are more likely to support authentic authorship and sustained learning (Bearman & Ajjawi, 2019; Boud & Soler, 2016, p. 401; Chaudhry & Kazim, 2022, p. 158).

The affective consequences are substantial. Students describe fear, anxiety, and reduced confidence, suggesting a psychosocial burden associated with algorithmically mediated assessment and reduced intellectual risk-taking under continuous scrutiny (Lim et al., 2023). Equity considerations further complicate detection practices, as neurodivergent students, multilingual writers, and those with distinctive styles appear disproportionately vulnerable to misclassification (Liang et al., 2023, p. 100779). Taken together, these findings suggest that generative AI exposes existing tensions in assessment design, institutional trust, and authorship expectations. Rather than indicating a deficit in student values, the findings point to a breakdown in relational trust between students and institutions. These findings support a shift away from detection-led governance towards design-led approaches that prioritise transparency, equity, and learning. Assessment

design could reduce reliance on surveillance by incorporating authentic tasks, oral or in-person assessment components, staged submissions, and process-oriented evaluation that foregrounds disciplinary reasoning over easily automated outputs. Rather than treating AI literacy as a student deficit, institutions should develop their own literacy regarding the emotional and surveillance toll of detection-oriented governance. The findings suggest students are already sophisticated navigators of AI tools and the systems that monitor them; the gap lies in institutional understanding of how detection frameworks generate anxiety, erode trust, and disproportionately penalise linguistic diversity. Institutions should also develop transparent and procedurally fair processes for addressing suspected AI misuse. Detection outputs ought to be treated as indicative, not determinative, supported by structured human review, clear appeal pathways, and consistent adjudication practices. Regular auditing of detection systems is required to identify and mitigate differential impacts, particularly for neurodivergent and multilingual students, ensuring that technological governance does not reproduce or amplify existing inequities.

## Conclusion

This study examined how awareness of AI academic integrity policies shapes student experience and ethical reasoning in higher education. The findings indicate that generative AI intensifies existing tensions in assessment design, institutional trust, and authorship expectations rather than functioning as a standalone disruption. Students navigate an assessment environment marked by uncertainty and risk, where they continuously balance authenticity, compliance, and academic survival. These findings support reconceptualising academic integrity as a relational and design-oriented construct grounded in transparency, fairness, and pedagogical support. Future research should incorporate interviews and cross-institutional sampling to extend and validate these findings, addressing the limitations of relying on publicly available social media data, which may over-represent heightened or polarised perspectives.

## References

Bearman, M., & Ajjawi, R. (2019). Can a rubric do more than be transparent? Invitation as a new metaphor for assessment criteria. *Studies in Higher Education, 46*, 1–1https://doi.org/10.1080/03075079.2019.1637842

Bearman, M., Tai, J., Ajjawi, R., Boud, D., & Dawson, P. (2026). Feeling AI: Circulating emotions, institutional climates, and moral boundaries in student use of AI. *Higher Education*. https://doi.org/10.1007/s10734-026-01658-6

Boud, D., & Soler, R. (2016). Sustainable assessment revisited. *Assessment & Evaluation in Higher Education, 41*(3), 400–413. https://doi.org/10.1080/02602938.2015.1018133

Chaudhry, M. A., & Kazim, E. (2022). Artificial intelligence in education (AIEd): A high-level academic and industry note 2021. *AI and Ethics, 2*(1), 157–165. https://doi.org/10.1007/s43681-021-00074-z

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.

Gerlich, M. (2025). AI tools in society: Impacts on cognitive offloading and the future of critical thinking. *Societies, 15*(1), 6. https://doi.org/10.3390/soc15010006

Liang, W., Yuksekgonul, M., Mao, Y., Wu, E., & Zou, J. (2023). GPT detectors are biased against non-native English writers. *Patterns, 4*(7), 100779. https://doi.org/10.1016/j.patter.2023.100779

Lim, T., Gottipati, S., & Cheong, M. (2025). What students really think: Unpacking AI ethics in educational assessments through a triadic framework. *International Journal of Educational Technology in Higher Education, 22*, 56. https://doi.org/10.1186/s41239-025-00556-8

Lim, W. M., Gunasekara, A., Pallant, J. L., Pallant, J. I., & Pechenkina, E. (2023). Generative AI and the future of education: Ragnarök or reformation? A paradoxical perspective from management educators. *The International Journal of Management Education, 21*(2), 100790. https://doi.org/10.1016/j.ijme.2023.100790

McGuirk, K. (2025). Navigating generative AI in higher education: A framework for educator decision-making. In Barker, S., Kelly, S., McInnes, R., & Dinmore, S. (Eds.), *Future Focussed. Educating in an era of continuous change. Proceedings ASCILITE 2025. Adelaide* (pp. 545-550). https://doi.org/10.65106/apubs.2025.2657

Moreno, M. A., Goniu, N., Moreno, P. S., & Diekema, D. (2013). Ethics of social media research: Common concerns and practical considerations. *Cyberpsychology, Behavior, and Social Networking, 16*(9), 708–713. https://doi.org/10.1089/cyber.2012.0334

Moya Figueroa, B., Eaton, S., Pethrick, H., Hayden, A., Brennan, R., Wiens, J., & McDermott, B. (2024). Academic integrity and artificial intelligence in higher education (HE) contexts: A rapid scoping review. *Canadian Perspectives on Academic Integrity, 7*. https://doi.org/10.55016/ojs/cpai.v7i3.78123

Perkins, M., Roe, J., Postma, D., McGaughran, J., & Hickerson, D. (2023). Detection of GPT-4 generated text in higher education: Combining academic judgement and software to identify generative AI tool misuse. *Journal of Academic Ethics, 22*, 1–25. https://doi.org/10.1007/s10805-023-09492-6

Risko, E. F., & Gilbert, S. J. (2016). Cognitive offloading. *Trends in Cognitive Sciences, 20*(9), 676–688.

Selwyn, N., Hillman, T., Eynon, R., Ferreira, G., Knox, J., Macgilchrist, F., & Sancho, J. M. (2020). What's next for Ed-Tech? Critical hopes and concerns for the 2020s. *Learning, Media and Technology, 45*, 1–6.



Shah, P., Patel, H., & Ladd, S. (2026). Caught in the algorithm: Student discourse on AI governance, surveillance, and academic integrity in higher education. In Cochrane, T., Narayan, V., Poole, T., Lai, P., Lake, P., Moore, O., & Li, L. (Eds.), *The Future is Connection: Collaboration, Creativity and Compassion in a Digital Society*. Proceedings ASCILITE 2026. Brisbane (pp. 808–813). https://doi.org/10.65106/apubs.2026.3369



Note: All published papers are refereed, having undergone a double-blind peer-review process.
The author(s) assign a Creative Commons by attribution license enabling others to distribute, remix, tweak, and build upon their work, even commercially, as long as credit is given to the author(s) for the original creation.

© Shah, P., Patel, H., & Ladd, S. 2026