The Algorithm's Benefit of the Doubt

Weekly Analysis — https://ainews.social

A student at Adelphi University sat down to defend something that should never have required a defense: the authorship of her own essay. She had written it. She said so. The university, armed with a detector's verdict, said otherwise, and the burden of proof landed not on the institution that made the accusation but on the student who had to disprove a number generated by software she had never seen, could not interrogate, and was not permitted to cross-examine [10]. The case has since become a lawsuit, one of a growing docket of them, and the pattern it crystallizes is the real subject of this essay: when a detection tool and a student disagree, the institution believes the tool.

[10] An Adelphi University student was accused of using AI to ... - Newsday

That reflex — granting the algorithm the benefit of the doubt that is withheld from the human being — is not a minor administrative quirk. It is an inversion of the entire logic of academic due process, and it has happened across higher education with remarkable speed and almost no public debate. The Adelphi suit is now tracked alongside dozens of others in a running tally of accusations that have curdled into litigation [4]. Each case rehearses the same scene. A tool flags. An instructor reports. A committee convenes. And somewhere in the proceedings the question that should be foundational — does this tool actually work? — goes unasked, because the institution has already decided that it does.

[4] AI Cheating Lawsuits Tracker — Every Case, Who Won (2026)

This essay is a survey of what is happening to higher education as AI arrives, but it approaches that vast question through a single revealing wound. The false-positive crisis is not a story about cheating students. It is a story about institutions that adopted a class of surveillance technology to solve a problem they never measured, granted that technology an evidentiary authority it never earned, and in doing so quietly rewrote the contract between the university and the people it is supposed to serve. To understand the landscape of stakeholder positions, the conflict between institutional priorities and pedagogical concerns, and the strange silence where a collaborative framing ought to be, you can do worse than follow the trail of a single flagged paragraph.

The Epidemic That Was Never Counted

Begin with the premise that justified the whole apparatus. Detection tools were sold into universities as the antidote to a cheating epidemic — a sudden, civilization-threatening surge of machine-written assignments that honest faculty could no longer distinguish from genuine work. The premise was plausible, and panic is a powerful procurement officer. But notice what was missing at the moment of adoption: evidence. Institutions did not first establish the base rate of AI-assisted cheating, then evaluate whether a given detector could identify it at an acceptable error rate, then weigh those errors against the cost of false accusation. They bought the tools and asked questions later, if at all.

The questions, when they finally came, were not reassuring. The University of Cambridge, examining whether large language models could even reliably *grade* student essays — a far easier task than catching covert authorship — found that the technology was "not yet good enough," that it rewarded surface fluency over substance, mistaking "style over" genuine quality [7]. If a model cannot be trusted to tell a good essay from a bad one, the confidence with which its cousin technology claims to tell a human essay from a machine one should collapse on contact. Yet that confidence is precisely what disciplinary committees have leaned on.

[7] AI not yet good enough to mark university essays, rewarding 'style over ...

The deeper problem is that detection inverts the relationship between a tool and the thing it claims to measure. As Meredith Broussard argued in [11], our culture is afflicted by "technochauvinism" — the faith that a computational solution is inherently superior to a human one, even when no evidence supports the claim. Her account of the post-2016 surprise among engineers, who could not understand why anyone believed what they read online, captures the same epistemics in reverse: institutions now believe what they read in a confidence score precisely because it is generated by a machine. The number *feels* like measurement. It performs objectivity. And the performance is enough to shift the burden of proof onto a teenager.

[11] Artificial Unintelligence

This is the first thing the sculpture professor down the hall needs to understand about the landscape: the cheating epidemic was real as a feeling and unverified as a fact. The most candid recent account of student behavior, an ethnographic study with the telling title "Everyone's using it, but no one is allowed to talk about it," describes a campus culture in which generative AI is ubiquitous, normalized, and driven underground precisely by the punitive posture of detection [1]. The tools meant to suppress AI use have instead made it unspeakable — which is a very different thing from making it rare. The epidemic

[1] "Everyone's using it, but no one is allowed to talk about it": College ...

was never counted because counting was never the point. The point was to be seen doing something.

Classification as an Act of Power

To see why detection went wrong, it helps to name what a detector actually is. It is a classifier. It sorts a population into two bins — "human" and "machine" — and attaches a probability to the sort. Kate Crawford, in [11], insists that we treat classification not as a neutral technical operation but as an exercise of authority: "Classification is an act of power," she writes, and embedded in working infrastructures these acts become "relatively invisible without losing any of their power." A confidence score of 98 percent is such an embedded classification. By the time it reaches a disciplinary hearing it has shed every trace of its uncertainty and presents itself as a finding of fact.

[11] The Atlas of AI

The power in that act is asymmetrical, and the asymmetry is the scandal. When the tool is right, the student is caught. When the tool is wrong, the student is still caught — and now must mount a defense against evidence that is structurally undefendable. The fullest legal analysis of this dynamic, from the Harvard Undergraduate Law Review, lays out the mechanism precisely: detection tools function as "opaque evidence" whose internal workings are proprietary, unauditable, and therefore impossible for an accused student to contest, threatening the most basic guarantees of due process [5]. You cannot cross-examine a number. You cannot subpoena a model's training data. You cannot ask the algorithm why it was so sure.

[5] AI Detection Tools and Academic Punishment: How Opaque Evidence ...

And the stochastic nature of the underlying technology makes the whole exercise more treacherous than its marketing admits. UNESCO's [11] warns that generative systems are "more likely to be stochastic in generating outputs or predictions, as the same inputs may lead to different outputs," rendering their content "potentially less trustworthy." The same warning applies to the detectors built atop the same statistical foundations: feed them the same essay twice and you may not get the same verdict. A system that cannot reproduce its own judgments is being used to render judgments that end careers.

[11] AI competency framework for teachers

Notice how rarely this technical reality enters the institutional conversation. The dominant frame remains compliance and risk management — how to write a policy, how to limit liability — rather than epistemology. Even the more sophisticated reporting on how universities are migrating "from the norm to the criterion," abandoning blanket rules for case-by-case judgment, treats the shift as a governance

achievement rather than an admission that the detection paradigm has failed on its own terms [13]. The instinct to manage the controversy is everywhere. The instinct to ask whether the tool deserves its authority is almost nowhere.

[13] De la norma al criterio: cómo las universidades encaran el ... - LinkedIn

The Chilling Effect on the Page

The most insidious cost of false positives is not borne by the students who are formally accused. It is borne by the much larger population who never face a hearing but who learn, from the surrounding atmosphere of suspicion, to write defensively. This is the chilling effect, and it strikes precisely at the cognitive behaviors a university exists to cultivate.

Consider what a student does once she internalizes that a detector is reading over her shoulder. She flattens her prose, because fluency is suspect. She avoids the sophisticated constructions that a classifier might misread as machine-generated. She keeps her drafts, screen-records her writing sessions, and times her keystrokes — not to learn, but to assemble an alibi. The Cambridge finding that detection-adjacent models reward "style over" substance has a perverse corollary: students now have an incentive to suppress style, to write worse on purpose, because polished writing reads as guilt [7]. The tool meant to protect the integrity of writing degrades the writing itself.

[7] AI not yet good enough to mark university essays, rewarding 'style over ...

The harm compounds with the genuine cognitive risks that AI use, separate from detection, already poses. A growing literature documents what researchers call "cognitive offloading" — the outsourcing of intellectual effort to a machine, producing a kind of "metacognitive laziness" in which students stop monitoring their own thinking [2]. This is a real pedagogical concern, and it deserves a serious pedagogical answer. But the detection regime answers it with surveillance rather than instruction, and surveillance produces its own offloading: the student offloads her sense of intellectual ownership to the apparatus that polices her. She no longer asks "is this good?" but "will this pass?" Both questions outsource judgment. Only one of them is the one the university claims to oppose.

[2] a1_Pereza_metacognitiva_y_descarga_cognitiva_en_ ...

The systematic evidence on whether AI functions as an "amplifier or substitute" for student cognition is genuinely mixed, which is itself the point: the research community has not settled the question that institutions have already answered with their procurement [9]. And the broader risk assessment is sobering. A widely covered report concluded that, as currently deployed, the risks of AI in schools outweigh

[9] Amplifier or substitute? A systematic review of generative ...

the benefits — not because the technology is worthless, but because the implementations are reckless [21]. The detection apparatus is the recklessness made concrete. It manufactures the very disengagement it was meant to prevent, teaching students that the safest relationship to their own writing is wary, minimal, and self-suspicious.

[21] Report: The risks of AI in schools outweigh the benefits : NPR

There is also a distributional cruelty here that the discourse barely registers. Detectors disproportionately misfire on non-native English writers, on neurodivergent students whose syntax is atypical, on anyone whose voice does not match the statistical center of the training data. The same surveillance logic that drives student-monitoring software — programs that, as the Associated Press has documented, carry serious risks even as they proliferate — falls hardest on the students least equipped to mount a defense [20]. The chilling effect, in other words, is not evenly distributed. It is coldest where the institution's protection is most needed.

[20] Programas de IA para monitorear a estudiantes tienen riesgos de ...

Governance Eats Pedagogy

Survey the discourse of the past year and one imbalance dominates everything: the conversation about AI in higher education is overwhelmingly a conversation about governance, and only marginally a conversation about learning. Read the trade press and you will find the genre crystallized in a single Forbes headline declaring that AI is now "fundable in higher ed — but only with real governance" [6]. The sentence is revealing precisely because it sounds responsible. Governance is the password that unlocks the money. The pedagogy is assumed to follow, and it rarely does.

[6] AI Is Now Fundable In Higher Ed—But Only With Real Governance - Forbes

The institutional how-to literature reinforces the frame. Guides on how "to build AI governance your school or university can actually use" treat the central challenge as one of policy architecture — committees, frameworks, acceptable-use documents — rather than as one of teaching and learning [18]. Surveys of generative AI policies at the world's top universities catalog an emerging consensus around disclosure and permission structures [17]. EDUCAUSE's authoritative state-of-play assessment maps an entire sector organizing itself around the management of AI rather than the integration of it into the actual work of forming minds [22]. This is not nothing. Policy matters. But policy has crowded out the prior question of what we are trying to teach and how a given tool helps or hinders it.

[18] How to build AI governance your school or university can ...

[17] Generative AI Policies at the World's Top Universities: 2026 ...

[22] The Current State of Play: AI in Higher Education and the Road Ahead

The conflict between institutional priorities and pedagogical concerns surfaces most visibly when an administration signs an enterprise deal over the objections of its own faculty. The California State Uni-

versity system's agreement to deploy ChatGPT across its campuses did exactly that, polarizing students and faculty who were largely presented with the arrangement as a fait accompli [12]. Here the governance-versus-pedagogy imbalance shows its true shape: governance, in practice, often means a procurement decision made at the top, with the pedagogical consequences delegated downward to the people who had no say in the purchase. The administrators who sign the contract bear the institutional risk. The faculty who teach the classes inherit the disruption. The students who pay tuition become, without consent, the user base.

[12] Cal State's deal for ChatGPT polarizes students and faculty - CalMatters

There is a workplace mirror to this dynamic that should give universities pause. The World Economic Forum's study of AI at work describes organizations discovering that the value of these tools depends far less on the technology than on how it is integrated into actual practice — that productivity hacks imposed from above tend to fail where genuine organizational transformation requires buy-in from the people doing the work [19]. Higher education is repeating the productivity-hack error at scale: it has treated AI adoption as an IT decision and a compliance exercise, when the only version that could possibly work is a slow, participatory, pedagogical one. The governance frame is not wrong. It is radically insufficient, and its sufficiency is being assumed.

[19] PDF AI at Work: From Productivity Hacks to Organizational Transformation

The Inversion of Due Process

Return now to the student in the hearing room, because the governance-pedagogy imbalance has a victim, and the victim is procedural justice. The deepest claim of this essay is that institutions have inverted due process — that they extend to a proprietary classifier the presumption of reliability they deny to a human being who says, simply and truthfully, "I wrote this."

Consider the ordinary architecture of fairness. In any defensible disciplinary system, the accuser bears the burden of proof; the accused can examine the evidence; the standard of proof scales with the severity of the consequence. The detection regime violates all three. The accuser is software whose error rate the institution has never independently established. The evidence is a confidence score whose derivation is a trade secret, which the Harvard analysis correctly identifies as the core threat: "opaque evidence" that no accused party can meaningfully contest [5]. And the standard of proof — what should be very high, given that an academic-integrity finding can end a degree and a career — is effectively set by a vendor's marketing claim about

[5] AI Detection Tools and Academic Punishment: How Opaque Evidence ...

accuracy.

The lawsuits are the predictable result of this inversion, and they are the mechanism by which the false-positive crisis converts into a crisis of institutional legitimacy. The growing docket tracked across the sector represents students refusing the inverted burden and forcing the question into a venue — the courtroom — where opaque evidence and unexamined accusers do not enjoy automatic deference [4]. The Adelphi litigation is emblematic: a student who maintains she wrote her own work, compelled to litigate against an institution that took a machine's word over hers [3]. Whatever the legal outcomes, the reputational outcome is already arriving. Each suit is a public advertisement that the university's word, when it accuses, may be worth less than the student's.

[4] AI Cheating Lawsuits Tracker — Every Case, Who Won (2026)

[3] Adelphi University accused a student of using AI to ... - Newsday

This is where the category-specific stakes become unavoidable. For an institution, a false positive is not a rounding error in an otherwise sound system. It is a breach of the academic contract — the implicit promise that the university will treat its students as honest until proven otherwise, through evidence they can see and answer. Every false accusation that survives appeal, and every true accusation built on undefendable evidence, withdraws a deposit from the institution's reserve of trust. The MIT Press primer [11] frames the governing question of automated decision-making with stark economy: "How many decisions and how much of those decisions do we want to delegate to AI? And who is responsible when something goes wrong?" The detection regime has answered the first question expansively and the second question not at all. When the tool is wrong, no one is responsible. The student absorbs the harm; the vendor disclaims liability; the institution hides behind the score.

[11] AI Ethics

An institution that cannot answer the question of responsibility has no business wielding the tool. The specificity bar — the rate of true negatives, the assurance that an innocent student will not be flagged — must be set extraordinarily high precisely because the cost of a false positive is the legitimacy of the whole disciplinary apparatus. No detector currently on the market clears that bar with public, independently audited evidence. The honest conclusion is that institutions should abandon tools that cannot meet it, rather than continuing to launder vendor confidence into disciplinary fact.

From Detection to Authentication

If detection is the failure, what is the alternative? The most promising answer reframes the entire problem. Instead of asking "did a ma-

chine write this?" — a forensic question the technology cannot reliably answer — institutions can ask "can this student demonstrate the thinking behind this work?" That is a pedagogical question, and the university is uniquely equipped to answer it. The shift is from detection to authentication, from policing outputs to designing assessments that make process visible.

This is not utopian. It is a return to forms the university already knows: the oral defense, the in-class write, the staged draft, the annotated bibliography, the conversation in which a student is asked to extend her own argument in real time. The emerging recognition that instructional design has become "the competence that AI makes decisive" points exactly here [15]. Assessment that is robust to AI is assessment that surfaces the human thinking AI cannot fake on demand — not because a detector caught the machine, but because the task itself requires what only the student can supply.

[15] El diseño instruccional, la competencia que la IA vuelve decisiva

The evidence that well-designed AI-mediated instruction can genuinely help, rather than merely surveil, is real and worth taking seriously. A randomized controlled trial published in *Nature* found that AI tutoring could outperform in-class active learning on certain measures — a result that should chasten anyone tempted to reject the technology wholesale [8]. Studies of AI feedback on student work likewise show measurable learning gains when the tool is positioned as a coach rather than a cop [14]. The point is not that AI has no place in the classroom. The point is that its defensible place is in the open, integrated into instruction by faculty who chose it, rather than imposed as a hidden tribunal that students must outrun.

[8] AI tutoring outperforms in-class active learning: an RCT ... - Nature

[14] Effects of Artificial Intelligence Feedback on Students ... - Springer

Authentication also dissolves the underground culture that detection created. Recall the ethnographic finding that everyone uses these tools while no one is permitted to discuss them [1]. A regime built on detection forces dishonesty by making honest disclosure dangerous. A regime built on authentication and disclosure — where students are asked to document how they used AI, and graded on the quality of that engagement — converts the same behavior from a punishable secret into a teachable practice. The emerging work on evaluating AI competency programs suggests the contours of what could be assessed instead: not whether the machine touched the work, but whether the student developed genuine judgment about when and how to use it [16].

[1] "Everyone's using it, but no one is allowed to talk about it": College ...

[16] Evaluation indicator system for AI certificate programs

What is conspicuously absent from almost the entire discourse is the partnership framing — the possibility that students might be co-designers of the policies that govern them rather than the objects those policies act upon. The Cal State controversy shows what hap-

pens in the partnership's absence: a top-down deal that polarizes the very people it was meant to serve [12]. The WEF's lesson from the workplace — that imposed productivity tools fail where participatory integration succeeds — is the lesson higher education has not yet learned [19]. The missing perspective in the AI-in-education conversation is the one in which the governed have a voice in the governance. That absence is not an oversight. It is the structural expression of the same inversion that lets a classifier outrank a student.

[12] Cal State's deal for ChatGPT polarizes students and faculty - CalMatters

[19] PDF AI at Work: From Productivity Hacks to Organizational Transformation

What the Flagged Paragraph Reveals

Step back from the single flagged paragraph and the whole landscape comes into focus. Higher education met the arrival of generative AI not with curiosity but with control, and it reached for control through a technology of suspicion before it had any evidence that the technology worked. It bought detectors to fight an epidemic it never counted, granted those detectors an evidentiary authority their stochastic foundations cannot support, and in the process taught a generation of students that their own words are guilty until a black box declares them innocent.

The cost is showing up on three ledgers at once. On the pedagogical ledger, the chilling effect flattens writing and corrodes the intellectual ownership the university exists to build, even as the genuine risk of cognitive offloading goes unaddressed by anything but surveillance [2]. On the institutional ledger, every false positive that survives appeal and every accusation built on opaque evidence withdraws trust the university cannot easily replace, while the lawsuits convert that erosion into public record [4]. On the legitimacy ledger — the deepest one — the institution has staked its credibility on tools that cannot meet a high specificity bar, and credibility, once spent on a vendor's confidence score, does not return on demand.

[2] a1_Pereza_metacognitiva_y_descarga_cognitiva_en_ ...

[4] AI Cheating Lawsuits Tracker — Every Case, Who Won (2026)

The way out is not more sophisticated detection. It is the abandonment of the detection paradigm in favor of authentication, transparency, and genuine partnership — assessment designed by faculty to surface human thinking, appeals processes that let students examine and contest the evidence against them, and governance that treats the governed as participants rather than suspects. The technology itself can be part of that future; the RCT and feedback evidence shows it has real pedagogical value when deployed in the open [8]. What cannot be part of it is the reflex this essay began with — the institutional instinct to believe the machine over the person.

[8] AI tutoring outperforms in-class active learning: an RCT ... - Nature

Broussard's diagnosis of technochauvinism in [11] is, in the end,

[11] Artificial Unintelligence

the most useful frame for what the past year has revealed. The university's error was not that it adopted a flawed tool; every institution adopts flawed tools. The error was that it gave the flawed tool the benefit of the doubt it owed its own students — that it mistook a confidence score for a finding of fact and a vendor's claim for evidence. Restoring the academic contract begins with reversing that gift. The benefit of the doubt belongs to the human being who says she wrote her own essay. It was never the algorithm's to keep.

References

1. "Everyone's using it, but no one is allowed to talk about it": College ...
2. a1_Pereza_metacognitiva_y_descarga_cognitiva_en_la_era_de_la_IA ...
3. Adelphi University accused a student of using AI to ... - Newsday
4. AI Cheating Lawsuits Tracker — Every Case, Who Won (2026)
5. AI Detection Tools and Academic Punishment: How Opaque Evidence ...
6. AI Is Now Fundable In Higher Ed—But Only With Real Governance - Forbes
7. AI not yet good enough to mark university essays, rewarding 'style over ...
8. AI tutoring outperforms in-class active learning: an RCT ... - Nature
9. Amplifier or substitute? A systematic review of generative ...
10. An Adelphi University student was accused of using AI to ... - Newsday
11. Artificial Unintelligence
12. Cal State's deal for ChatGPT polarizes students and faculty - CalMatters
13. De la norma al criterio: cómo las universidades encaran el ... - LinkedIn
14. Effects of Artificial Intelligence Feedback on Students ... - Springer
15. El diseño instruccional, la competencia que la IA vuelve decisiva

16. Evaluation indicator system for AI certificate programs

17. Generative AI Policies at the World's Top Universities: 2026 ...

18. How to build AI governance your school or university can ...

19. PDF AI at Work: From Productivity Hacks to Organizational Transformation

20. Programas de IA para monitorear a estudiantes tienen riesgos de ...

21. Report: The risks of AI in schools outweigh the benefits : NPR

22. The Current State of Play: AI in Higher Education and the Road Ahead