The New Sorting Hat:

AI Writing Detection as Classification Infrastructure

Jake Lawrence

Independent Researcher

Author Note

Jake Lawrence is an independent researcher based in Illinois. This paper applies Susan Leigh Star and Geoffrey Bowker's classification-as-infrastructure framework to AI writing detection in higher education. It is the companion paper to The Sorting Machine: Special Education Classification as Infrastructure, which applies the same theoretical lens to IDEA disability categories.

Correspondence concerning this paper should be addressed to Jake Lawrence. Email: jake@jakelawrence.xyz

Abstract

This paper argues that AI writing detection tools should be understood not as technology products undergoing evaluation but as classification infrastructure being installed across higher education in real time. Drawing on Bowker and Star's (1999) theory of classification systems and Star's (1999) concept of infrastructure, the analysis identifies three structural problems that infrastructure theory predicts will compound rather than self-correct as the system becomes more embedded: a boundary problem (the conversion of continuous probability scores into binary categories), a looping effect (students and faculty adapting their behavior to the detection system, degrading the practices the system claims to protect), and a pattern of disparate impact that tracks existing demographic lines in educational inequality. The paper further argues that the human/AI binary on which these tools depend is a category fallacy: an incoherent classification imposed on a continuum of human-machine collaboration. Three policy proposals are advanced: mandatory demographic transparency in false positive reporting, sunset clauses in institutional detection contracts, and a due process floor prohibiting the use of detection scores as the sole basis for academic integrity proceedings. The analysis draws on recent federal litigation, peer-reviewed studies of detection bias, and the detection companies' own documentation to demonstrate that the infrastructure's failures are structural, not incidental, and that the installed base of institutional dependencies will resist correction through technical improvement alone.

Keywords: AI writing detection, classification infrastructure, Turnitin, academic integrity, disparate impact, infrastructure theory, looping effects, higher education policy

The New Sorting Hat:

AI Writing Detection as Classification Infrastructure

In February 2025, a federal judge in the Eastern District of New York found that Adelphi University's decision to punish a student for AI plagiarism was "without valid basis and devoid of reason" (Newby v. Adelphi University, 2025). Orion Newby, a freshman with documented learning differences who had worked with university-provided tutors throughout his writing process, submitted an essay on Christianity and Islam. His professor ran it through Turnitin's AI detector. The tool flagged it as wholly AI-generated. Newby denied using AI.

The school gave him a zero and ordered him into an anti-plagiarism course, a remediation program for a category he had been wrongly sorted into, then warned that a second offense could mean suspension. When he appealed, the school declined to reexamine the allegations. It never interviewed his tutors. It never examined his drafts. It had a score. Newby's family spent six figures on legal fees to prove what a conversation could have established for free.

His case is not unusual. Only the outcome is. At Yale, a French-born MBA student is suing after GPTZero flagged his exam answers; the university suspended him for a year, and his lawsuit alleges discrimination against non-native English speakers. At the University of Michigan, a student with OCD and generalized anxiety disorder is suing after professors interpreted her formal, structured writing style as evidence of AI generation. Across campuses nationwide, students describe running their own work through multiple detectors before submitting, trying to pre-clear assignments they wrote themselves.

The public conversation about these tools has mostly been a technology debate: Do they work? Are they accurate enough? This paper argues that these questions miss the structural problem entirely. AI writing detectors are not technology products being evaluated by educators. They are classification infrastructure being installed across education in real time (Bowker & Star, 1999; Star, 1999), and infrastructure theory predicts that their deepest failures will compound, not self-correct, as the system becomes more embedded.

Theoretical Framework: Classification as Infrastructure

This analysis draws on the infrastructure studies tradition, particularly Bowker and Star's (1999) theory of classification systems and Star's (1999) ethnography of infrastructure. In this framework, infrastructure is not simply the thing underneath the thing. It is a system that gets woven into institutional processes so thoroughly that it becomes invisible: taken for granted, difficult to question, and nearly impossible to remove. Infrastructure persists not because it is optimal but because the cost of replacing it exceeds the cost of living with its flaws.

Star (1999) identified eight properties of infrastructure, three of which are central to the present analysis. Embeddedness: infrastructure is sunk into other structures, social arrangements, and technologies; changing one node requires renegotiating the entire network. Installed base: infrastructure inherits strengths and limitations from the base it was built on; it does not grow de novo. Visibility upon breakdown: infrastructure is transparent when working and becomes visible only when it fails.

Hacking's (1995) concept of looping effects provides a complementary lens: when a classification system is applied to people, the people classified change their behavior in response, which in turn changes what the classification captures. The classification and the classified co-construct each other.

The present paper applies this framework to AI writing detection, arguing that these tools exhibit each of Star's three properties and that Hacking's looping effect is already visible in student and faculty behavioral adaptation.

AI Detection as Embedded Infrastructure

Over 16,000 institutions are currently integrated with Turnitin's AI detection system. Detection scores feed into grading workflows, academic integrity databases, and student disciplinary records. Faculty build syllabi around the assumption that submitted work will be scanned. Policies reference detection thresholds. The tool hooks into Canvas, Blackboard, and Moodle, running automatically on submission.

The depth of this embeddedness was illustrated in April 2023, when Purdue University announced that Turnitin's AI detection update would activate automatically and disclosed that the university "does not have the ability to turn it off" (Purdue University, 2023). The institution that licensed the tool could not control when the tool changed underneath them. That is not a product. That is infrastructure.

The cascade of dependencies runs from the detection tool through LMS integration, grading workflows, academic integrity databases, syllabus policies, and pedagogical choices (including, increasingly, the replacement of take-home essays with in-class handwritten exams, not because they are pedagogically superior, but because they are scan-proof). Removing the tool would leave five downstream systems wired to expect its output. That is what infrastructure means: everything gets built on top of it, and eventually, removing the foundation means demolishing the building.

Three Structural Problems

The Boundary Problem

AI detectors produce continuous probability scores: a percentage likelihood that text was generated by a machine (Elkhatat et al., 2023). But institutions need categories: cheating or not cheating, flagged or clean. So they impose thresholds. Where the line is drawn determines who gets accused.

Turnitin acknowledged this problem seven weeks after launch. Their Chief Product Officer admitted the company had discovered “a higher incidence of false positives” when less than 20% of AI writing is detected (Turnitin, 2023a). She did not disclose the actual rate. Their fix was to display an asterisk instead of a number for scores between 1% and 19%. They did not improve the tool in that range. They stopped showing the score. The infrastructure kept running; they dimmed the dashboard light.

Turnitin advertises a document-level false positive rate of less than 1%. But faculty do not experience the document-level abstraction. They open a report and see highlighted sentences. The sentence-level false positive rate is approximately 4% (Weber-Wulff et al., 2023). At that rate, a twenty-sentence paper has a 56% probability of containing at least one falsely flagged sentence. The number the company advertises and the number the user experiences are measuring different things.

Table 1

Illustrative Detection Scores Across Student Profiles

Profile	Score	Actual AI use	At 20% threshold

Native English, creative style	3%	None	Clear
ADHD, structured writing	12%	None	Clear
French-born, MBA program	18%	None	Clear
ESL, first-generation	24%	None	False positive
Brainstormed with AI, wrote by hand	41%	Partial	Flagged
OCD, meticulous revision	67%	None	False positive
Pasted prompt, submitted output	89%	Full	Flagged
Generated and submitted unchanged	95%	Full	Flagged

Note. Scores are illustrative, based on patterns reported in the literature. Profiles are composites drawn from published case descriptions and litigation records.

The Looping Effect

Students who know they will be scanned change their writing (Hacking, 1995). Not to write better, but to write in ways that avoid triggering detection. They write less clearly, less concisely, less precisely, because clarity and predictability are exactly what the detectors flag.

Turnitin's own documentation explains that it evaluates how statistically likely each word choice is given its context: language models choose optimal words, so human writers need to choose suboptimal ones to avoid suspicion (Sadasivan et al., 2023). Better writing now looks more "AI-like" to the machine.

The loop runs continuously. Detection is deployed; students adapt their writing; writing quality degrades; AI "humanizer" tools emerge (over 150 of them, charging up to $50/month, drawing 33.9 million website visits in a single month; SimilarWeb, 2024); detectors retrain to catch new evasion patterns, shifting the threshold again. Faculty adapt in parallel: in-class handwritten exams replace take-home essays, not because they are pedagogically superior but because they are scan-proof. The classification system is restructuring pedagogy around itself.

Students are paying for the privilege of making their own writing worse so that a machine will believe they wrote it. The system is producing exactly the adversarial behavior it was designed to prevent, then pointing to that behavior as justification for its own existence.

Disparate Impact

Liang et al. (2023) tested seven AI detectors and found that they classified over 61% of TOEFL essays written by non-native English speakers as AI-generated, while performing near-perfectly with essays by U.S.-born students. The reason is structural: non-native speakers tend toward simpler syntax, more predictable vocabulary, and more formulaic constructions. These are the exact features the detectors associate with machine output.

Table 2

Disparate Impact in AI Detection False Positive Rates

Population	False positive rate	Source
TOEFL essays (non-native speakers)	61.3%	Liang et al. (2023)
U.S.-born student essays	3.2%	Liang et al. (2023)
Black students	~20%	Liang et al. (2023)
White students	~7%	Liang et al. (2023)

Note. Rates are drawn from the seven-detector comparison study. Approximate values (~) indicate estimates across multiple detectors.

Students with autism, ADHD, and other conditions that produce systematic or highly structured writing patterns face elevated false positive rates (Weber-Wulff et al., 2023). The detectors are not failing randomly. They are failing along the same lines that educational systems have always failed along. And calling the result a technical measurement.

The Category Fallacy

Beneath the three structural problems lies a deeper one: the category itself is incoherent. The detectors assume a binary, human-written or AI-generated, that is already obsolete.

Consider six students submitting the same assignment. One asked an LLM to brainstorm topics, then researched and wrote the essay herself (22% AI probability). One drafted from scratch, asked an LLM to identify weak arguments, and revised by hand (31%). One dictated rough ideas, fed the transcript to an LLM for an outline, and wrote from the outline (38%). One wrote a full draft, asked the LLM to suggest transitions, rejected most, and adopted two (19%). One asked the LLM to generate a first draft, then substantially rewrote every paragraph in her own voice (55%). One pasted in a prompt and submitted the output unchanged (91%). Five of six students learned something. The detector cannot see any of that. It can only measure resemblance.

The detector cannot distinguish between the student who used AI to think and the student who used AI to avoid thinking. What is being built is a system that punishes students for how their writing sounds rather than how they learned.

Kleinman (1988) called this the “category fallacy”: importing a classification scheme into a context where it does not fit. The human/AI binary made a kind of sense before AI became a writing environment rather than a writing tool. It does not make sense now. But the infrastructure has already been built on top of it.

The Infrastructure Knows Its Own Limits

The company that built the infrastructure knows it (Turnitin, 2023b). The following disclaimer appears at the bottom of every Turnitin AI detection report:

> Our AI writing detection model may not always be accurate... so it should not be used as the sole basis for adverse actions against a student.

Turnitin has published guides for educators on how to handle false positive conversations. They have published separate guides for students on what to do when falsely accused. They have built an entire secondary infrastructure: professional development materials, conversation protocols, remediation workflows. All to manage the consequences of the primary infrastructure’s failures.

This is the signature of mature infrastructure. The system’s operators can describe its limitations in detail, publish documentation acknowledging those limitations, and build an entire support apparatus for managing the consequences of those limitations, all while continuing to sell and expand the system. The disclaimer does not constrain the infrastructure. The disclaimer is part of it.

Predictions

Infrastructure theory predicts what happens next. The tools will become invisible in the technical sense: not hidden, but taken for granted. Unquestioned. That is what infrastructure scholars mean by “transparent” (Star, 1999): not that you can see through it, but that you stop seeing it at all.

Switching costs will compound. Every policy, workflow, and database that references a detection score makes the system harder to remove. Disparate impacts will accumulate. False flags travel in student records, shaping institutional trajectories long after the detection score itself is forgotten. The binary will collapse. Writing is already becoming collaborative in ways no detector can parse. But the infrastructure will persist because everything has been built on top of it.

The diagnostic categories in the DSM have persisted for decades despite widespread acknowledgment of their scientific limitations, not because they are accurate but because insurance billing, treatment guidelines, and clinical training are all built on top of them. AI detection is on the same trajectory, compressed from decades into years because digital infrastructure installs faster than institutional infrastructure.

Three Policy Proposals

What would it mean to treat AI detection as an infrastructure design problem rather than a technology procurement decision? At minimum, three things.

Proposal 1: Demographic Transparency

If a detection tool cannot report its false positive rate by race, by language background, and by neurodiversity status, it should not be deployed in any context where those false positives carry consequences. This is not a feature request. It is a precondition for responsible deployment.

Proposal 2: Sunset Clauses

Every AI detection contract should require re-evaluation on a fixed schedule (two years, not perpetuity) so institutions are forced to actively decide whether the infrastructure still serves them rather than allowing it to persist by default. Without mandatory re-evaluation, switching costs compound every semester: more policies written, more workflows built, more training delivered. Within three years, removing the tool becomes more expensive than keeping it, regardless of whether it works.

Proposal 3: A Due Process Floor

An absolute prohibition on using detection scores as the sole basis for academic integrity proceedings. Turnitin's own documentation already says this, in fine print, on every report, beneath the score that institutions are already treating as a verdict. The next Orion Newby will not have a family that can afford federal litigation. That student will accept the zero, the remediation course, and the record.

Conclusion

Orion Newby got his grade back and his record expunged. It cost his family six figures and a federal court order (Newby v. Adelphi University, 2025). The infrastructure will sort the next student tomorrow morning, before anyone reads the disclaimer at the bottom of the screen.

The argument of this paper is not that AI writing detection tools are inaccurate, although they are. It is not that they are biased, although they are. It is that they are infrastructure: embedded in institutional processes, built on an installed base of dependencies, and invisible to the people operating inside them until they break down. The accuracy debate treats these tools as instruments that can be calibrated. Infrastructure theory predicts that they will be absorbed, normalized, and maintained, not because they are good enough, but because the cost of removing them will exceed the cost of living with their failures.

The question is not whether to improve the tools. The question is whether institutions will recognize what they have installed before the installed base makes that recognition irrelevant.

References

Bowker, G. C., & Star, S. L. (1999). Sorting things out: Classification and its consequences. MIT Press.

Elkhatat, A. M., Elsaid, K., & Almeer, S. (2023). Evaluating the efficacy of AI content detection tools in differentiating between human and AI-generated text. International Journal for Educational Integrity, 19(17). https://doi.org/10.1007/s40979-023-00140-5

Hacking, I. (1995). The looping effects of human kinds. In D. Sperber, D. Premack, & A. J. Premack (Eds.), Causal cognition: A multidisciplinary debate (pp. 351–394). Oxford University Press.

Kleinman, A. (1988). Rethinking psychiatry: From cultural category to personal experience. Free Press.

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

Newby v. Adelphi University (2025). Decision and order granting preliminary injunction. U.S. District Court, Eastern District of New York.

Purdue University. (2023). Turnitin AI detection: What instructors need to know. Purdue Online Writing Lab announcement.

Sadasivan, V. S., Kumar, A., Balasubramanian, S., Wang, W., & Feizi, S. (2023). Can AI-generated text be reliably detected? arXiv preprint arXiv:2303.11156. https://doi.org/10.48550/arXiv.2303.11156

SimilarWeb. (2024). AI humanizer tool traffic analysis. SimilarWeb Digital Intelligence.

Star, S. L. (1999). The ethnography of infrastructure. American Behavioral Scientist, 43(3), 377–391. https://doi.org/10.1177/00027649921955326

Turnitin. (2023a). Understanding false positives within our AI writing detection capabilities. Turnitin Blog. https://www.turnitin.com/blog/understanding-false-positives-within-our-ai-writing-detection-capabilities

Turnitin. (2023b). AI writing detection: Educator FAQ and resource guide. Turnitin Resources. https://www.turnitin.com/solutions/ai-writing

Weber-Wulff, D., Anohina-Naumeca, A., Bjelobaba, S., et al. (2023). Testing of detection tools for AI-generated text. International Journal for Educational Integrity, 19(26). https://doi.org/10.1007/s40979-023-00146-z