The Detector's Dilemma: Why False Positives Are an Architectural Limit

Weekly Analysis — https://ainews.social

There is a particular kind of product that sells you the cure and the disease in the same breath. AI writing detectors are such a product. They arrived to solve a panic that generative text had created, promising to restore a distinction—human versus machine—that the technology itself had dissolved. The pitch is seductive in its symmetry: if one algorithm can write like a person, another can surely catch it. But the symmetry is false, and the falseness is not a bug to be patched in the next release. It is built into the floor. When Vanderbilt University quietly switched off Turnitin's AI detection feature in 2023, it did not cite a temporary glitch or a pricing dispute; it cited the tool's inability to tell the truth reliably enough to act on, and the disproportionate damage done when it was wrong [4].

[4] Guidance on AI Detection and Why We're Disabling Turnitin's AI Detector

That decision is worth dwelling on, because it inverts the usual order of technology adoption. Normally a tool is embraced on the strength of vendor claims and abandoned only after years of disappointing field data. Here, an institution with every incentive to want the tool to work—a real problem, a paying contract, an anxious user base—looked at the mechanism and concluded that the mechanism could not be fixed. The interesting question this raises is not whether detectors are *currently* inaccurate. Everyone, including the vendors, concedes some error rate. The interesting question is whether the error is structural: whether false positives are an accident of immature engineering that better data and bigger models will eventually retire, or whether they are an architectural limit, a permanent feature of what these instruments are and how they work.

I want to argue the latter, and I want to argue it on the evidence rather than the intuition. The case against detectors is usually made on moral grounds—it feels unjust to accuse a person of cheating on the word of a black box. That is true, but it is also easy to wave away as squeamishness. The stronger case is technical and economic. Detectors model "human writing" as a statistical silhouette derived from narrow training data; that silhouette can be spoofed by anyone who reads a single how-to thread, and it systematically misreads writers who never fit the silhouette in the first place. The result is a category of tool whose failures fall hardest on the people least able to contest

them, and whose marketing has run conspicuously ahead of its science. To see why, you have to look at what the detector actually does, as opposed to what it says it does.

What the Detector Claims to See

Strip the marketing away and an AI text detector is a classifier. It ingests a passage and emits a probability—this text is X percent likely to be machine-generated—on the basis of statistical properties it has learned to associate with model output. The two properties that do most of the work are *perplexity*, a measure of how surprising each next word is given the preceding ones, and *burstiness*, the variation in sentence length and structure across a passage. Large language models, optimized to predict the likeliest next token, tend to produce text that is low in perplexity and smooth in burstiness: fluent, even, statistically unsurprising. Human writing, the theory goes, is lumpier. The detector hunts for the smoothness.

The first thing to notice is that this is a claim about *averages*, dressed up as a claim about *individuals*. A vendor can demonstrate, on a curated benchmark, that machine text scores lower perplexity than human text on the whole. But the tool is not deployed to judge populations; it is pointed at a single document written by a single person, and asked to render a verdict on that one case. The Spanish-language review of the research literature on AI plagiarism detection is blunt about the consequence: under realistic conditions the tools are far less reliable than their dashboards suggest, and their accuracy collapses precisely when the stakes are individual rather than aggregate [10]. A statistic about distributions is being asked to function as a fingerprint about persons, and it cannot bear the weight.

[10] ¿Funcionan los detectores de plagio con IA? Esto es lo que dice la investigación

The second thing to notice is that the ”human” the detector is trained to recognize is not humanity in general. It is a specific, historically situated sample of text—whatever corpus the vendor assembled—and that corpus encodes a particular register: standard, fluent, idiomatic, usually edited, usually written by people for whom the prestige dialect of the training language comes naturally. This is where the abstract problem of perplexity becomes a concrete problem of justice. A writer whose prose is unusually clean and even—because they have learned the language formally, because they write to a template they were taught, because they are neurodivergent and favor regular structure, because English is their third language and they reach for safe constructions—produces exactly the low-burstiness signature the detector reads as synthetic. The tool does not detect machine writing.

It detects *deviation from the writing of the people in its training set*, and then mislabels that deviation as guilt.

This is not a novel pathology of detectors; it is the standard pathology of machine classification, and the literature on it is old enough to be embarrassing. Janelle Shane's account of how learning systems absorb the priorities of their data is mordant on exactly this point: a model trained on a skewed sample does not learn the world, it learns the skew, and it will reproduce that skew with mechanical confidence while presenting it as objectivity [9]. Kate Crawford makes the harder structural version of the argument in her dismantling of emotion-recognition systems, which claim to read interior states from surfaces and largely do not; the detector belongs to the same family of instruments that "do not do what they claim," tools that convert a contestable inference into an authoritative-looking readout [9]. The detector's perplexity score has the same epistemic status as the emotion engine's "anger: 87%." It is a number that looks like a measurement and behaves like a guess.

[9] You Look Like a Thing and I Love You

[9] The Atlas of AI

The Fingerprint That Was Never There

The deepest problem is that the signal the detector hunts for is not stable, and it was never going to be. Two forces are eroding it simultaneously, and both move faster than any vendor can retrain.

The first is convergence. Detectors assume a durable stylistic gap between human and machine prose, but that gap is closing from both sides. Each model generation is tuned to sound more human—more varied, more idiosyncratic, less robotically smooth—which is to say each generation is explicitly optimized to defeat the very feature detectors rely on. Meanwhile, an enormous and growing share of ordinary human writing is now drafted, edited, or polished with the assistance of the same models, so the "human" baseline is itself migrating toward the machine signature. Stanford's index of the field documents how quickly capability and ubiquity have advanced on both fronts, with model performance and adoption climbing in lockstep [7]. A detector calibrated on last year's gap is, by construction, mis-calibrated on this year's. The target it was built to hit is receding faster than the instrument can be re-aimed.

[7] The 2026 AI Index Report - Stanford HAI

The second force is that the gap, even where it persists, carries no causal information. A low-perplexity passage is consistent with a machine having written it. It is equally consistent with a careful human writer, a non-native writer using cautious constructions, a student writing to a rubric, or a person who simply writes plainly. The

detector cannot distinguish among these because the underlying signal does not distinguish among them. The 2024 edition of Stanford's responsible-AI survey already foregrounded the empirical studies showing detector outputs to be unreliable across exactly these confounds, a finding that has only hardened since [9]. What the detector measures is real—some texts genuinely are smoother than others—but the inference it draws from that measurement is unsupported. It is reading a correlation and reporting a cause.

[9] HAI_AI-Index-Report-2024

This is why "improve the accuracy" is not a coherent roadmap. You can lower the false-positive rate, but only by raising the threshold at which the tool cries machine, which means letting more actual machine text through—the false negatives climb as the false positives fall. The two errors trade against each other along a curve fixed by how much real signal exists, and the real signal is thin and shrinking. There is no setting on the dial that gives you a low false-positive rate *and* a high catch rate, because the information needed to have both is not present in the text. A careful adopter should treat any vendor claim of "99% accuracy with near-zero false positives" not as an impressive result but as a category error or a benchmark artifact—a number generated under conditions that do not resemble the messy, adversarial, heterogeneous reality in which the tool is actually used. The proofreader's guide to false positives, written from inside the detection-adjacent industry, concedes as much when it walks through how routinely clean human writing trips the alarm [3].

[3] False Positives in AI Detection: Complete Guide 2026

How Trivially the Signal Breaks

If the underlying signal were merely thin, that would be damning enough. What makes the architectural verdict unavoidable is how easily the signal can be erased on purpose. Adversarial robustness—the ability of a classifier to hold up when someone is actively trying to fool it—is the property detectors most need and most conspicuously lack.

The defeats are not exotic. They do not require a research lab or an obfuscation toolkit. A person who wants machine-generated text to read as human can ask the model to vary its sentence lengths, to write in a less formal register, to insert a few idiosyncrasies, or simply to "rewrite this so it doesn't sound like AI." Light paraphrasing, a pass through a second model, the substitution of a handful of synonyms, the manual roughening of a paragraph or two—each of these perturbs the statistical surface enough to push a passage back across the threshold. The signal the detector depends on is fragile in the pre-

cise sense that it can be removed by the same conversational interface that produced the text, with no special skill, in seconds.

There is a bleak comedy in where this leads, and the French-language coverage captured it sharply: in order to prove they had *not* used AI, students began learning to use AI more skillfully—to launder their own legitimate work into a form the detector would accept, or to disguise machine output well enough to pass [5]. The tool, deployed to suppress a behavior, ends up teaching it. This is the signature of an instrument fighting the wrong battle: every countermeasure it introduces becomes a tutorial for its own evasion.

[5] Le paradoxe des détecteurs d'IA en classe : pour prouver qu'ils ne trichent pas en se servant de l'IA, des étudiants apprennent à utiliser l'IA

Shane's framing of adversarial attacks is useful here precisely because she refuses to treat them as a special category. The fake stop sign, the cartoon tunnel painted on a rock wall—machine perception is vulnerable to deception in the same way human perception is, because perception works by pattern and patterns can be forged [9]. The detector is a pattern-matcher, and its pattern is public, cheap to study, and trivial to spoof. An entire cottage industry of "humanizer" services now exists for no purpose other than to launder machine text past these classifiers, which means the detector is not merely beatable but is being beaten at scale, continuously, by the very population it is meant to police. A security tool that loses to its adversary by default is not a security tool. It is theater.

[9] You Look Like a Thing and I Love You

And the asymmetry of that theater cuts the wrong way. The motivated bad actor—the person actually determined to pass off machine writing as their own—is the one most likely to learn the evasions and slip through clean. The person who gets caught is disproportionately the one who *wasn't* gaming anything: the honest writer whose natural prose happens to sit in the danger zone, who has no idea the danger zone exists, and who therefore takes no steps to avoid it. The tool inverts its own purpose. It waves the guilty through and detains the innocent, because the guilty are studying the lock and the innocent are not.

The Economics of a Broken Instrument

If the engineering case were the whole story, detectors would have quietly died, the way bad products usually do. They have not, and the reason is economic. There is a market here—a real, funded, contracted market—and the market's incentives are misaligned with the truth in instructive ways.

Consider the procurement data. An investigation into what institutions are actually paying found more than fifteen million dollars

committed to AI detection services, a figure that represents not experimentation but infrastructure, multi-year contracts wired into the plumbing of how written work gets processed [8]. That spending creates a constituency. Once an institution has paid for a detector, integrated it, and trained staff to read its outputs, it acquires a sunk-cost interest in believing the tool works. The vendor, for its part, has every reason to keep the accuracy claims buoyant and the error-rate disclosures buried in fine print. This is the ordinary physics of a platform that has achieved lock-in: the question shifts from "does this work?" to "how do we justify what we've already bought?"

[8] What Universities Spend on AI Detection — $15M+ in Data

The deeper point is about framing, and it generalizes beyond detectors to the whole AI-tools market. The dominant way these products are presented—roughly a quarter of all tools coverage—is the neutral language of *utility*: a detector is simply a tool, an instrument, a feature you toggle on, morally weightless, like a spell-checker. That framing is doing enormous covert work. It positions the detector as a passive lens that merely reports what is already there, when in fact the detector is an active accuser that manufactures a verdict and assigns it a number. The "it's just a tool" register launders a probabilistic guess into an institutional fact. Microsoft's own adoption-tracking apparatus illustrates how thoroughly the industry has normalized this utility framing, scoring organizations on how deeply AI features penetrate their workflows as though penetration were itself the measure of value [1]. When adoption is the metric, the question of whether the adopted thing actually does what it claims recedes from view entirely.

[1] AI Adoption Category in Adoption Score

This is where the skeptical adopter has to be most disciplined, because the tools market is not evenly balanced between enthusiasm and doubt. Coverage skews toward the promotional—the case study, the success story, the productivity testimonial—and the genre conventions of that material are designed to foreground capability and background failure. The vendor case-study format is practically a literary genre of its own, a sequence of frictionless wins with the error bars edited out [6]. Against that current, the documented failures arrive as interruptions rather than as the baseline they should be. The honest accounts of what happens when a detector is wrong—the accused writer with no recourse, the appeals process that presumes the machine over the person—exist, but they are not what the market is built to surface [2]. A careful reader has to actively correct for the slope of the available evidence, weighting the rare skeptical account more heavily precisely because the system is built to produce fewer of them.

[6] Power Platform and Copilot Studio real-world case studies

[2] AI Detection False Positives: Real Stories, Real Consequences

The Asymmetry of a False Accusation

Every classifier makes two kinds of error, and a mature engineering culture asks which kind is more expensive before deploying. For a detector, the two errors are wildly unequal in cost, and the architecture systematically loads the cost onto the party least equipped to bear it.

A false negative—machine text waved through as human—costs little in any individual instance. Some work goes ungraded for its true provenance; the world turns. A false positive—human text condemned as machine—costs a great deal: an accusation of dishonesty, a damaged record, a process of appeal in which the accused must prove a negative against a number they cannot inspect. The Spanish-language synthesis of the research is explicit that this asymmetry is not incidental but central; the harm of the false positive is categorically greater than the harm of the false negative, which means a tool that trades catch rate for false-positive rate is trading the cheap error for the expensive one in exactly the wrong direction [10]. And because the false positives concentrate among non-standard writers—second-language users, the formally trained, the neurodivergent, the plain stylist—the expensive error is also a discriminatory one. The tool does not distribute its mistakes randomly. It distributes them down the existing gradient of linguistic privilege.

[10] ¿Funcionan los detectores de plagio con IA? Esto es lo que dice la investigación

This is the precise structure that critical scholarship on automated systems has been describing for years: a technology marketed as a neutral arbiter, which in operation encodes and amplifies the disadvantages of the people it judges, while wrapping that judgment in the unappealable authority of the algorithm. Crawford's account of systems that claim to read truth off surfaces and instead project the assumptions of their builders maps onto the detector with almost no translation; the perplexity score, like the emotion-recognition readout, is an inference masquerading as an observation, and the masquerade is the whole problem [9]. When Vanderbilt disabled Turnitin's detector, the reasoning was exactly this calculus of asymmetric harm: the cost of being wrong, multiplied by the certainty of being wrong sometimes, exceeded any plausible benefit of being right [4].

[9] The Atlas of AI

[4] Guidance on AI Detection and Why We're Disabling Turnitin's AI Detector

There is a further, subtler harm that the utility framing conceals: the chilling effect on the writers themselves. A person who knows that clean, even prose triggers the alarm faces a perverse incentive to write worse—to introduce noise, to roughen their sentences, to abandon the clarity they were taught to prize—in order to read as authentically human to a machine. The detector does not merely punish; it deforms. It exerts a quiet pressure on writing to conform to whatever the classifier thinks a person sounds like, which is to say it

makes the model's narrow image of humanity into a behavioral target. This is the tool *doing something to you*, in the most direct sense: reaching past the document and shaping the writer's hand.

Provenance, or Nothing

If detection is architecturally unsuited to its task, the question becomes what, if anything, can do the job it was hired for. The honest answer is that the job may not be doable from the text alone—and that is the real lesson, the one the vendors are most reluctant to state.

The structural reason detection fails is that it tries to recover provenance from a finished artifact that carries no reliable trace of its origin. Once text exists, the question "who or what produced this?" is, in the general case, unanswerable from the text itself, because the same string of words could have arrived by a dozen routes. This is why the only technically serious alternatives move *upstream*, to the moment of generation, where provenance can be inscribed rather than inferred. Watermarking—embedding a statistical signature into model output at the point of creation, a signature that can later be checked—is the leading candidate, and its appeal is precisely that it does not depend on guessing from surface features. It knows because it was there. The trajectory of the field's own self-assessment points this way: the serious work on machine-text provenance has shifted from post-hoc classification toward generation-time marking, an acknowledgment that the detector's whole epistemic stance was backwards [7].

[7] The 2026 AI Index Report - Stanford HAI

But watermarking is not a clean salvation, and the skeptical reader should resist swapping one piece of vendor optimism for another. A watermark only works if the model that wrote the text chose to embed one, which means it does nothing about open-weight models a user runs locally, nothing about systems whose makers decline to participate, and not much about a determined adversary who paraphrases the marked text through an unmarked model and washes the signature out. The same adversarial fragility that doomed detection threatens watermarking too, just one layer up. What watermarking offers is not certainty but a *changed default*: instead of presuming guilt from a statistical silhouette and forcing the accused to disprove it, a provenance system can offer positive evidence of machine origin where the evidence genuinely exists, and remain silent where it does not. That is an architecturally honest posture in a way detection never was. It claims to know only what it can actually know.

The mature position, then, is not "detection needs more accuracy" but "detection was the wrong category." This is the maturity crisis

the tools face, and it does not resolve into a comfortable upgrade path. A vendor can pivot toward genuinely explainable, context-aware, provenance-grounded systems that report their own uncertainty and decline to manufacture verdicts they cannot support—or a vendor can keep selling the silhouette and accept that it is selling theater. There is no third option in which the perplexity classifier quietly becomes reliable, because the unreliability is not in the engineering. It is in the premise.

A careful adopter walks away from this evidence with a short, sharp checklist of questions, and they are not the questions the marketing invites. Not "how accurate is it?"—accuracy on a curated benchmark is nearly meaningless—but "what is your false-positive rate on non-native and neurodivergent writers, measured in the field?" Not "does it catch AI?" but "how is it defeated, and how cheaply?" Not "can I see the score?" but "can I see *why*, in terms a human can audit and an accused person can contest?" The detector that cannot answer these is not a tool with room to improve. It is an instrument whose central claim—that it can read origin off a surface—is false at the level of physics, and no amount of model scale will make a false thing true. Shane's rule for machine learning was to expect the bias and design for it [9]; the detector's makers did neither, and the people paying the price for that omission are the ones whose only crime was to write in a voice the machine had never learned to recognize as human.

[9] You Look Like a Thing and I Love You

References

1. AI Adoption Category in Adoption Score
2. AI Detection False Positives: Real Stories, Real Consequences
3. False Positives in AI Detection: Complete Guide 2026
4. Guidance on AI Detection and Why We're Disabling Turnitin's AI Detector
5. Le paradoxe des détecteurs d'IA en classe : pour prouver qu'ils ne trichent pas en se servant de l'IA, des étudiants apprennent à utiliser l'IA
6. Power Platform and Copilot Studio real-world case studies
7. The 2026 AI Index Report - Stanford HAI
8. What Universities Spend on AI Detection — $15M+ in Data
9. You Look Like a Thing and I Love You

10. ¿Funcionan los detectores de plagio con IA? Esto es lo que dice la investigación