

On (not) using AI detectors

Martin Paul Eve

Professor of Literature, Technology and Publishing
Birkbeck, University of London

Associate Director of Knowledge Commons
Michigan State University

31 August 2026



Image: AI Paper Mills. On the left is a grey-scale version of a rat with inaccurate and oversized reproductive organs recognisable from an academic article which contained AI slop. The image has been edited to be sliced, and blue painted torus icons overlay the image. The background features a green mountainous range with a magenta tiled floor and gradient sky. (Marcin Wilkowski / <https://betterimagesofai.org/> / <https://creativecommons.org/licenses/by/4.0/>)

Cite as: Martin Paul Eve, "On (not) using AI detectors", *eve.gd: Martin Paul Eve*, August 31, 2026, <https://doi.org/10.59348/hxwam-6hk75>.

DOI: <https://doi.org/10.59348/hxwam-6hk75>

Available at: <https://eve.gd/2026/08/31/on-not-using-ai-detectors/>

A frequent discussion that I have had with an academic colleague, but also with various team members at Knowledge Commons and also with those at OLH, pertains to the use of “AI writing detectors” to unearth cheating students and/or dishonest academic researchers.

By way of background, I agree with [UCB Law’s policy on AI](#) and the taxonomy they have use for LLM-based activities. They prohibit – and I agree with this ruling – AI use in: conceptualizing, outlining, drafting, revising, translating, and editing.

Brave New Regulatory World

The most recent impetus for this discussion was in light of new EU legislation that [requires companies to identify AI output](#). Basically, any publisher, anywhere, where the material might be accessed in the EU, *must* disclose any AI use unless it has definitely been through a process of human review. So, if a lax editor at an OLH journal uses AI to review a piece(!) and then publishes it, OLH could be legally exposed.

This would be extremely beneficial as it might means I waste less time reading something only to find, halfway through, that nobody invested any time in it and it is AI slop. But enforcement of this, at a publisher or platform, is going to require a huge amount of labour: a dedicated person (or team at some places) before long. At OLH, for example, there is [a thorough AI policy](#). But to enforce this and to ensure that editors do not in any way violate the condition that we must have a human reviewer (e.g. making sure reviewers are NOT sneakily using AI to help them review and respond to papers) requires constant oversight. This is expensive stuff.

This regulation was relatively quick to emerge (for the law). It had to be, because otherwise everyone says “the regulation can’t keep up”. And I do really think that the web would be a much better place if all AI content had to be labelled as such. Especially on educational or research matters. And legislation is probably the only way to achieve that, although it too may not succeed.

It also only applies to “text which is published with the purpose of informing the public on matters of public interest”. So, I joked, that rules out most humanities research (*drumroll, cymbal crash, badum tish*). (*I am allowed to make this joke as I am a humanities professor and also clearly do not really believe it*. Indeed, Eve ([2014](#), *et passim*) has consistently argued that there is a public appetite for humanities research, shown by the large number of people who choose to study such subjects at university every year.)

So the publisher world is facing new interesting regulatory need to be able to detect AI-generated prose. So what can we do...

The AI Detectors: You Are Using AI If You Use Them

There are a range of tools that can be used to detect AI/LLM prose. But the first thing I want to say is that *all of these tools are themselves AI*. If you are someone who radically objects to AI in every circumstance and thinks that it needs to be burned to the ground, then by using these tools *you are also using an AI system*.

Now my view: I am very sceptical of these tools. AI cannot detect AI more reliably than any other system, unless it knows the model weights used to generate the text (and the input parameters). [Pangram](#), one such system, does seem to do a marginally better job than other detectors right now from what I can see, but it’s all a game of probabilities and catch-up. (Except for the [AI systems’ own text watermarking](#), which allows *them* to sell a service that can probabilistically identify their own prose with much greater accuracy. It’s clever stuff, actually.)

(In fairness to Pangram, I had a run in with them a while back in fact, and [I was wrong](#).

This was over the Granta essay scandal where they checked past stories for AI. I was worried that the past stories would already be in their training data so it would not be a fair assessment. I was actually wrong (and fine/happy to admit it). They were good-natured in their reply.)

The Stakes Are High But Does The Accuracy Match?

OK, but for me, the issue with AI detectors, particularly in the student work (or, actually, the publishing) context, all comes back to this: let's consider that an AI detector has **95% true positive accuracy** with only a **5% false positive rate**. Sounds pretty convincing! I have no idea whether those numbers are realistic. Indeed, a recent amusing viral demonstration of one such tool showed a dire instance of AI generation in a text. That text began “**Call me Ishmael.**”

However, what these accuracy statistics *mean* depend on how you act on them. If anybody found to be cheating in a class by using AI is to be kicked out of the class, and you have 50 students, you *could* have to expel, *wrongly*, up to 3 or so students. Suddenly, this is far more worrying. Likewise, if you are going to block someone from publishing based on it, this is giving serious power to a machine classifier.

I have had the debate that such numbers can be used only as a guide, a sort of rough feel for whether or not a submission *might* be a piece of AI slop. But what does that “guide” actually mean? Surely it means that you have been subconsciously prejudiced by an AI machine that wrongly identifies *Moby-Dick* as the output of another machine model? If we use these tools to inform our judgments of whether a piece is AI generated or not, we *will*, at some point, undoubtedly, wrongly suspect a student or researcher, based on a black-box machine's finger pointing.

I *would* trust the watermarking system of the company that produced the AI model in the first place. But this means that you have to check with all the AI providers as to whether any of them generated it. And it is possible that local, specifically trained LLMs will not have such watermarking and cannot be detected.

The [Committee on Publication Ethics \(COPE\) also solicited advice](#), last year, in 2025, on the use of these tools:

The current status of AI detection software means that it is not sensible to apply a threshold approach. Sometimes text written by a human can be flagged as produced by AI if it uses very specific language and phrases, and AI indicators are still inconsistent enough that their output cannot be relied upon; they can both under-predict and over-predict AI usage. Most policies on generative AI are based on how the tool is used, how the output is verified, how transparent the authors are, and the editorial assessment rather than a certain threshold of acceptability.

Disability and Detection

One colleague, during this discussion, astutely noted a fear that false positives could flag marginalised communities more often because of their potentially different verbal patterns, thereby once more embedding discrimination in AI detectors.

Curiously, I wondered whether there could be an inverse effect. The detectors are looking for predictable, common patterns that would have been produced by an LLM, with specific (estimated) weights and generation parameters. If a speech pattern is statistically novel, and thereby under-represented in the training corpus (and so less likely to appear) etc., then surely it will be *less likely* to be the output of an LLM? That is, more novel speech that is seen less frequently is less likely to be detected as AI.

Indeed, it is notable that AI detectors are usually the inverse of the other types of text-generation AI, in their goals. The detectors want to find traces of the machine, while the generators want to make it invisible (watermarks notwithstanding).

What's The Conclusion?

A colleague pointed out, when discussing this, an article titled “[Careful With Those Accusations](#)” that details an actual use of the OLH AI policy at work. Very good.

But anyway, here's where I end up on this:

1. If you use an AI detector, you are using AI. If you don't want to use AI at all, do not use AI detectors.
2. AI detectors are not accurate, they are probabilistic. Even the watermarking implemented by the companies themselves.
3. If you use their judgements bluntly, you will wrongly penalise someone, at some point. You will also wrongly exonerate others.
4. If you use their judgements merely as "guides", you will, at some time, also wrongly suspect someone, while you will incorrectly mentally exonerate some offenders. Even if you use this merely as a prompt, it will alter your subconscious appraisal of work, even when the machine could be *totally wrong* (call me Ishmael). Sometimes, it might be right, for sure. But other times it *will* be wrong. And you can never know in which camp you are falling.

I do not have a better plan for detecting this stuff. It is a huge problem for academia. The optimism of the "design new assessments that are AI-proof" crowd does not warm my heart. Essay-writing was such a huge developmental part of my intellectual trajectory. For a generation or more to lose that because machines can produce plausible (but often not even great) facsimiles of argumentative reason and expression of "thought" is, in my view, an educational tragedy.

I do not even like the idea of a return to closed-book or in-person examinations. I fared poorly in my school exams... *until I was allowed to type my answers on a computer*. I did not have internet or anything when writing, but the second I did not have to write by hand, my grades radically improved. Furthermore, actually doing research, as an undergraduate, working towards an essay, was the part of the training that I needed and wanted. Not my ability to sit in an exam room having magically learned in advance what I should have read to answer an argumentative/analytical question.

Featured Image

Credit to [Marcin Wilkowski](#) / [AI paper mills](#) / [Licensed under CC BY-4.0](#)

Thanks

Thanks for discussion on this go to JB, SE, SD, AB, MSL, CE and SH.

References

This reference list has been generated, in many cases, through an automated process and may therefore contain errors or incomplete metadata. It is provided purely as assistance for finding the items mentioned within this post.

‘AI Policy’, Open Library of Humanities <<https://www.openlibhums.org/site/ai-policy/>>

‘Artificial Intelligence Policy’, UC Berkeley Law

<<https://www.law.berkeley.edu/academics/registrar/academic-rules/artificial-intelligence-policy/>>

Eve, Martin Paul, *Open Access and the Humanities*, Cambridge University Press, 26 November 2014 <<https://doi.org/10.1017/CBO9781316161012>>

Eve, Martin Paul, ‘An interesting response here from Pangram; it turns out their system did not have peripheral knowledge of the Granta essays that were fed into it’, *Bluesky*, 23 May 2026 <<https://bsky.app/profile/eve.gd/post/3mmjoebggm225>>

Handling an article produced by AI, Committee on Publication Ethics (COPE), 13 May 2025 <<https://publicationethics.org/guidance/case/handling-article-produced-ai>>

Padolsey, James, ‘How AI text watermarking works: a visual guide’, *declaude*, August 2026 <<https://declaude.org/watermarking/>>

Pangram <<https://www.pangram.com/>>

‘The Stamp of Shame: What Claude’s New Legal Watermark Means for your Research’, *The Lex Academic Blog*, Lex Academic, 11 August 2026 <<https://www.lexacademic.com/blog/the-stamp-of-shame-what-claude-new-legal-watermark-means-for-your-research/>>

Weinberg, Justin, ‘Careful with those Accusations’, *Daily Nous*, 27 August 2026 <<https://dailynous.com/2026/08/27/careful-with-those-accusations/>>