0% Human

Two obligations went into the law. One of them was cheap to build.

THOMAS ROCHA III
AUG 18, 2026

On August 14, Google announced that users can turn off the visible watermark on AI-generated images, video, and music. The little sparkle in the corner is now a setting called Media Watermark. The invisible SynthID layer stays, the C2PA provenance metadata stays, and neither of those has a switch of any kind. The company was direct about why it made the change: removing the visible mark was one of the most requested fixes users asked for.

Every decision in the chain is locally reasonable.

Article 50 of the EU AI Act contains two obligations that people keep collapsing into one. The first says that if a system interacts directly with a person, the person has to be informed they are dealing with AI. The second says synthetic output must be machine-readable and detectable as synthetic. The first is addressed to a human being, in the moment, while something is happening. The second is addressed to an instrument, afterward, holding a detector.

Only one of those is buildable at the model layer.

You cannot put a declaration inside a token stream. A declaration is an act performed at a boundary, to the parties present, in a form they can perceive. What you can put inside a token stream is a statistical signature that survives some editing and not other editing, and that is what got built. Anthropic embeds an invisible mark in generated text across its products. The company says plainly that you will not see it, that it does not change the reading, that a detected mark is a signal rather than proof, that it can appear on text the model did not write, and that editing or translation or brevity can defeat it. That is an honest description of a forensic instrument. It is not a disclosure to anyone.

The visible sparkle was not a declaration either. It was a marker on the artifact, not an act performed toward a person. But it was the only layer in the entire stack that a human being could perceive without an instrument, and in practice it did the job the first obligation cares about: someone looking at the thing knew what they were looking at. It was also ugly, it was in the way, it made a client deck look amateurish, and it made the product look worse against competitors who never had one.

It is now optional, except in countries whose own law requires it to stay. Which tells you how the requirement is understood everywhere else.

Google is not lying to anyone, and Anthropic is not trying to take credit for your writing. The obligation that was cheap to satisfy is the one that got satisfied, and the obligation that was expensive, unpleasant, and product-degrading is the one that quietly became a setting. That is not corruption. That is what an equilibrium looks like when enforcement cannot exceed capability.

Think about what this leaves you with in the case that actually matters.

You are on a phone call. The voice is synthetic. Somewhere in the audio pipeline, a machine-readable provenance signal may well exist. You cannot hear it. You have no detector, and there is no artifact to run one against, because there is no file. There is only sound that has already passed. By the time anything could be inspected, the conversation is over, and you have already decided whether to trust what you were told. The layer that would have told you at the start was the one made optional.

The marking regime works beautifully for the case where someone comes back later with a captured file and asks what produced it. It does nothing at all for the case where you are in the room.

Then there is the part that made me laugh out loud.

I wrote an essay this week about authority in multi-agent systems. Before publishing, Substack ran it through a detection service. The result came back: 100% AI, 0% AI-assisted, 0% human.

That is not a provenance finding. It does not say a model was involved, or that the text shows machine influence. It reports a percentage of human authorship, and the percentage is zero. That is an authorship judgment, made from evidence incapable of observing authorship.

Here is what the zero is measuring. It is measuring the tokens. It is a statement about the surface of the text, produced by an instrument

built to answer one question, which is whether a language model was involved in generating this string. On that question, it is not wrong. A model was involved.

Here is what it is not measuring. Who chose the thesis. Who put two findings next to each other that nobody had connected. Who noticed the mathematical claim in the fourth section was overstated and cut it. Who decided a strong counterexample should be conceded in the author's own text before a critic could raise it. Who rejected two of a reviewer's four suggested edits because they would have made the piece safer and worse. Who cut every sentence that endorsed somebody else's mechanism. Who said no, twenty times, to things that would have been easier to keep.

None of that is in the tokens. All of it is the essay.

And consider what a real 0% human would have to look like. A system conceives the thesis, selects the argument, does the research, writes the draft, makes the graphic, picks the tags, and posts it to my account without my approval. That is the scenario the number describes. That is the rogue author the label implies. It is close to the exact inverse of what happened, and the detector gave the machine full credit for a job it did not do.

I wrote something in April about using a language model as scaffolding during a period when my own continuity was unreliable. The line I keep coming back to is about boundaries: inputs are opinions, analysis, or drafts; decisions are votes; and only one vote matters. We draft together. I press send.

A detector cannot see the votes. It was never built to. It can tell you that a machine was in the room, and it cannot tell you who was in charge, and those have never been the same question.

So here is my actual complaint, and it is not with any of these companies.

The law asked for two things. The market built the one it could build at the layer where it was cheap. And now the thing that got built is being received as though it answered the other one, because it is the only thing on offer and because a compliance artifact that exists always looks more like an answer than a gap that does not.

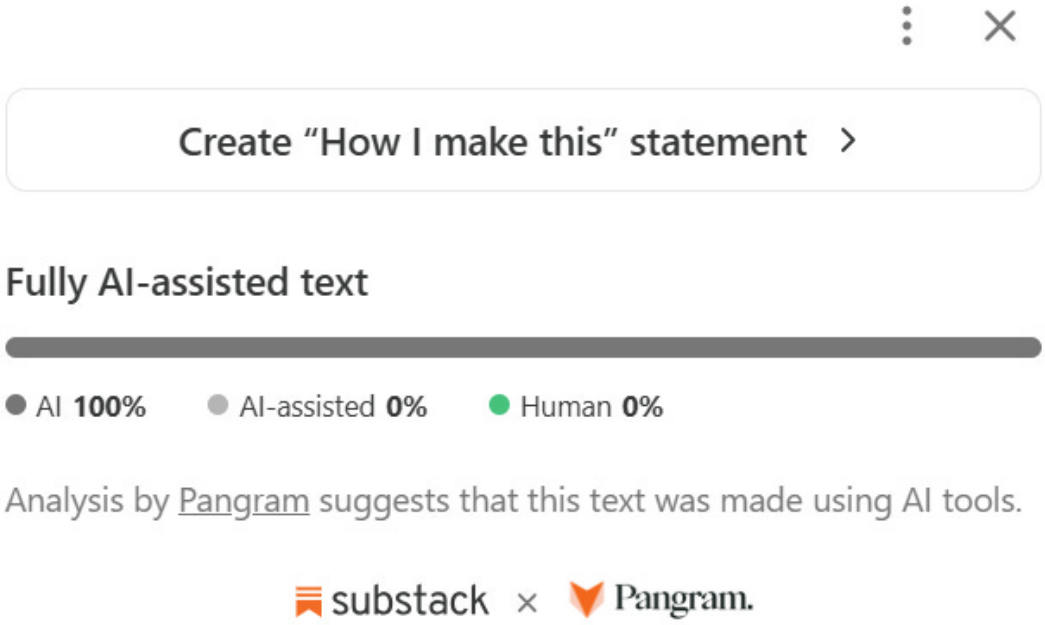
That is worse than the gap was. A gap generates pressure. A filled checkbox over a gap removes it. Before August, the honest position was that no general mechanism existed for proving, across a live interaction, who was told what, when, and under whose authority. Now there is a mark, a detector, a percentage, a settings toggle, and the appearance of a system that handles this. Anyone arguing for real declaration now has to dislodge a compliance instrument with regulatory endorsement instead of arguing into an admitted void.

The instruments we have measure contact. Something passed through here. That is worth knowing.

It is not the same as knowing who authorized what, or who was told what, or who decided. Those questions have to be answered at the boundary, while the thing is happening, by something other than the participant whose conduct is in question. We have not built that. We have built a very good way of finding fingerprints afterward, and we are calling it a doorbell.

My essay scored zero percent, human. I made every decision in it.

Both of those are true, and only one of them is measurable, and we are about to spend a decade governing on the one that is


Create “How I make this” statement
Fully AI-assisted text
AI 100%
AI-assisted 0%
Human 0%
Analysis by Pangram suggests that this text was made using AI tools.
substack × Pangram.