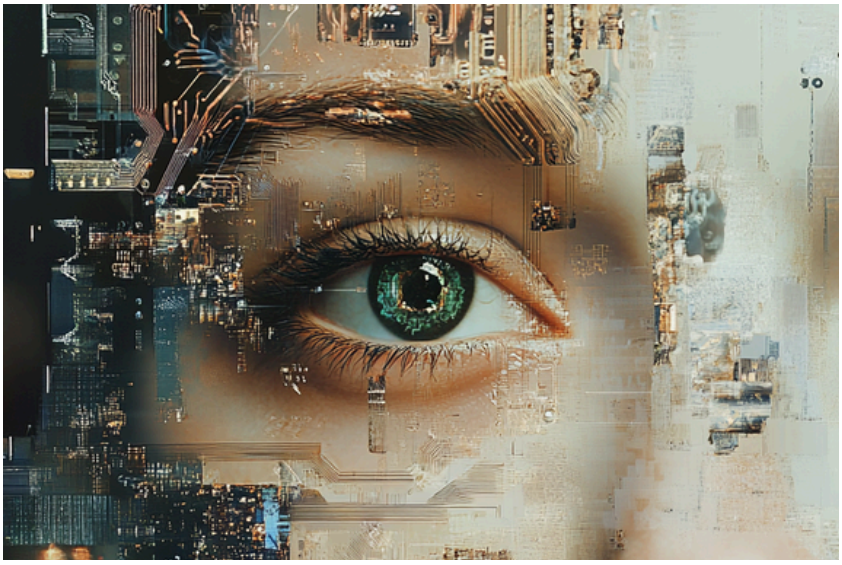
What Anthropic's New Watermark Actually Means Under the EU AI Act

Anthropic recently confirmed that Claude models launched from August 2, 2026, onward embed an invisible watermark directly into the text they generate. The update has been widely reported, and not always accurately.

UROŠ RAJIĆ

SENIOR ASSOCIATE



Some coverage described it as a breakthrough in detection. Others called it a privacy risk. Neither framing is quite right, and the difference matters for anyone building a product on Claude or using it to produce content that reaches EU users, including businesses based in Serbia and the wider Balkans.

What the watermark actually is

According to Anthropic's own documentation, the mark is woven into the generated text itself, not attached as metadata or hidden characters. It does not change the meaning, quality, or readability of the response. Because it lives inside the text, it can survive copying and pasting and may persist through some editing. It applies at the model level, so it shows up no matter which Claude product or surface the text comes from, whether that is the API, Claude.ai, Claude Code, or a third-party product built on top of the model. Anthropic is rolling this out worldwide, not only for content generated for EU users, so businesses in Serbia and the Balkans using Claude for non-EU markets will still see the mark on their output, even though the obligation that prompted it is an EU one.

Anthropic pairs this with a second technique for files. When Claude generates a supported file type, such as an SVG, PNG, or JPG, it attaches signed provenance metadata that follows the C2PA standard, the same open standard used across the industry to record where a piece of content came from.

Both mechanisms exist to support one specific legal commitment. Anthropic has signed the EU AI Act's Code of Practice on Transparency of AI-Generated Content, and this is Anthropic putting that commitment into working code.

Where this sits in the EU AI Act

Article 50 of the AI Act is easy to underestimate because it does not carry the heavy conformity assessments that apply to high-risk systems. But its reach is broader than almost any other provision in the Act, because it applies to any organization that falls into one of four situations, regardless of whether their AI use is high-risk at all.

Those four situations are worth naming plainly, because they fall on different actors in the chain:

- A provider whose AI system communicates directly with people, such as a chatbot, must design it so users know they are dealing with AI.
- A provider of a system that generates synthetic audio, image, video, or text, including general-purpose models like Claude, must mark that output in a machine-readable format that is detectable as AI-generated. This is the obligation Anthropic's watermark responds to.
- A deployer using AI for emotion recognition or biometric categorization must inform the people exposed to it.
- A deployer who publishes a deepfake, or AI-generated text meant to inform the public on a matter of public interest, must disclose that the content is AI-generated, unless a human has reviewed it and taken editorial responsibility for it.

The Code of Practice that Anthropic signed is a voluntary instrument built to operationalize these obligations, particularly the marking duty in the second bullet above. The European Commission and the AI Board have since assessed the Code as adequate for demonstrating compliance (though The Commission is careful to note that adherence to the Code is not, on its own, conclusive proof of compliance), which gives signatories a fair amount of legal certainty about how regulators will judge their marketing practices. Worth noting for anyone tracking the calendar: only one part of Article 50 actually moved under the AI Omnibus package. The machine-readable marking duty in Article 50(2) was postponed from August 2, 2026, to December 2, 2026, and only applies to

generative AI systems already on the market before August 2, 2026. The deployer disclosure duties under Article 50(1), (3), and (4) were not postponed at all and took effect on August 2, 2026, as originally scheduled. Anthropic chose to build marking into its models on the original timeline anyway.

What the mark does not do

This is the part that tends to get lost in the headlines, and it is the part that matters most for anyone relying on this as a compliance tool. Anthropic's own documentation is candid about the limits, and they are worth repeating without softening them.

A detected watermark is a signal, not proof. It tells you that Claude may have processed the content. It does not tell you that Claude wrote it from scratch. Someone might have used Claude only to proofread, translate, or summarize a document that a human otherwise wrote, and the output would still carry the mark. Content can also be edited or recombined after Claude processed it, which changes what the mark is actually vouching for.

The reverse is equally true. The absence of a detected watermark does not mean the content is human-written. Heavy rewriting can strip the signal. Short passages may not carry enough text for reliable detection. Screenshots and file format conversions routinely strip metadata entirely. And any model released before the marking rollout carries nothing at all, since Anthropic is still working through retrofitting older models and has not committed to a specific product timeline for doing so – though the regulatory backstop is December 2, 2026, the Omnibus deadline by which providers of pre-existing generative AI systems must add machine-readable marking.

None of this makes the watermark pointless. It is a genuine, useful signal for platforms and researchers building detection tooling. But it is compliance infrastructure, not proof of authorship, and it should not be treated as either in a contract, a dispute, or a regulatory filing.

Why does this matter even though Serbia is not in the EU

The AI Act's reach is not limited to companies established in the EU. It applies to providers and deployers outside the Union whenever the output of their AI system is placed on the EU market or used within it. A Belgrade-based company running an AI-powered customer chatbot for EU users, generating marketing content aimed at EU customers, or producing news-style content that gets published to an EU audience, can fall squarely within Article 50's deployer obligations even though it has no EU establishment at all.

This is where the distinction between what Anthropic does and what its customers must do becomes practically important. Anthropic's watermark satisfies Anthropic's own provider obligation under Article 50(2). It does not satisfy your obligation as a deployer under Article 50(1), (3), or (4) if your use of Claude falls into one of those categories. Those are separate, visible, human-facing disclosure duties, and building on a model that quietly marks its output in the background does none of that work for you.

What this means in practice

For any business building a product on the Claude API or using it to generate customer-facing content, three things are worth checking now rather than after a regulator asks.

First, work out which Article 50 category applies to your use case, since a chatbot, a content-generation tool, and a biometric feature each trigger a different disclosure duty and a different design requirement.

Second, do not assume the underlying model's watermark discharges your obligation if your product requires a visible AI disclosure or a deepfake label; these must be built into your product, not inherited from the model.

Third, keep your own documentation of how AI-generated content moves through your organization, since the watermark's limitations mean it will not reliably answer authorship questions on its own if a dispute or an audit arises later.

Worth flagging on the calendar alongside these three checks: February 2, 2027, is the deadline under the Code of Practice for providers to have an interoperability solution for watermark detection in place. That doesn't create a new duty for deployers. Still, it does mean the "signal, not proof" limitations discussed above should be somewhat more reliable – and worth revisiting – once that detection tooling lands.

Article 50 rewards organizations that treat transparency as a design requirement rather than an afterthought, and the deadlines are close enough now that this is worth a proper look rather than a guess.

GECIĆ | LAW
BELGRADE | BRUSSELS
WWW.GECICLAW.COM