

AI or Real? A Step-by-Step Guide to Detecting AI-Generated Images and Videos

Assem Hijazi – Experience, Knowledge & Ideas

Contents

- AI or Real? A Step-by-Step Guide to Detecting AI-Generated Images and Videos..... 1
- Introduction 3
- Step 1: Check for a Content Credential (C2PA) 4
 - What C2PA actually is..... 4
 - Who's actually using it in 2026 4
 - How to check it yourself 5
 - The limitations — read this part carefully 5
 - How to check it yourself 5
- Step 2: Scan for an Invisible Watermark 7
 - How watermarking differs from C2PA 7
 - Where this stands in 2026..... 7
 - The limitations — and this one matters a lot 7
 - "How to check for it" section for Step 2:..... 8
- Step 3: Run It Through an AI Detection Tool 9
 - How these tools actually work 9
 - Tools worth knowing, by use case..... 9
 - The accuracy reality check — don't skip this 10
 - Best practice: don't rely on one tool 10
- Step 4: Visual and Audio Inspection..... 11
 - The myth to drop: blinking..... 11
 - What still holds up in 2026..... 11
 - Audio-specific cues 12
 - If it's a live video call, not a recording..... 12
 - Optional forensic tools for this layer 12
 - The honest limitation of this whole step..... 13
- Step 5: Verify the Source 14

Reverse image and video search	14
Professional-grade tools (for higher-stakes verification)	14
What to actually check once you find the source	15
Practical checks:.....	15
The honest limitation here	15
Step 6: Weigh All the Signals Together	16
A simple framework for combining signals.....	16
How much rigor you actually need depends on the stakes.....	16
A word on false accusations	17
The bottom line	17

A few years ago, spotting an AI-generated photo was easy — too many fingers, melted text, waxy skin. That era is over. Today's generators produce images and video that can fool a careful eye in seconds of viewing, and that's exactly why "is this AI?" has become one of the most common questions people ask about anything they see online.

The honest answer is that there's no single foolproof test — no app you can point at a file that returns a certain "yes" or "no." What actually works is a layered process: checking provenance data, scanning for watermarks, running the file through detection tools, inspecting it visually, and verifying its source — then weighing all of that together. This guide walks through that process step by step.

Introduction

Every day, more of what we see online was never captured by a camera at all. A politician "caught" saying something outrageous, a product demo that looks too polished, a viral clip of a disaster that never happened — AI-generated images and video have gotten good enough that the old tricks (count the fingers, check for melted text) no longer cut it.

That leaves a real gap for anyone trying to make sense of what's real: journalists verifying footage before publishing, hiring managers screening video interviews, shoppers checking product photos, or just someone deciding whether to share a video with friends. The instinct is to look for "the one tool" that gives a clean yes-or-no answer. But that tool doesn't exist — and treating any single check as final is how people get fooled.

What actually works is closer to how a fraud investigator builds a case: no single piece of evidence is proof, but several independent signals pointing the same direction give you real confidence. This guide breaks that process into six steps, walking from the fastest checks to the most involved ones.

The 6-Step Verification Process

Step	What You Do	What It Tells You
1. Check for a Content Credential (C2PA)	Look for signed provenance data embedded in the original file	If present and valid, shows exactly which tool created it and what edits were made
2. Scan for an invisible watermark	Check for embedded signals like Google's SynthID	Can survive cropping/compression that strips metadata; confirms known AI tools
3. Run it through an AI detection tool	Use a pixel- or frame-level analyzer (Sightengine, Hive, TrueMedia, etc.)	Gives a probability score based on statistical artifacts, even with no metadata
4. Inspect it visually and audibly	Look for blinking patterns, lip-sync, hands, lighting, background physics	Reveals artifacts that generators still struggle to fully hide

5. Verify the source	Reverse-search the image or a video frame, check who posted it first	Confirms whether the content has a credible origin and history
6. Weigh all signals together	Compare results across steps 1–5	Builds overall confidence — no single step is decisive on its own

You won't always need to complete all six steps — often the picture becomes clear by step 2 or 3. But when the stakes are high (financial decisions, news reporting, identity verification), working through the full sequence is worth the extra few minutes.

Step 1: Check for a Content Credential (C2PA)

The single biggest shift in AI detection over the past year is that most major generators now automatically attach signed proof of their own origin to the files they produce. This is done through a standard called **C2PA (Coalition for Content Provenance and Authenticity)**, often surfaced to users as "Content Credentials."

What C2PA actually is

C2PA was built by a coalition including Adobe, Microsoft, Google, OpenAI, the BBC, and the New York Times. The coalition launched in 2021 as a merger of two earlier efforts: Adobe's Content Authenticity Initiative, which had been building image-provenance tooling since 2019, and Project Origin, a collaboration between the BBC, CBC, Microsoft, and the New York Times focused on news-media provenance. As of 2026, the standard has matured into an official ISO specification, giving it far more weight as an industry-wide format rather than a single company's feature.

A **Content Credential** is a signed manifest embedded in the file itself. It can record the device that captured an image, the software that edited it, the timestamp, the GPS location, and whether AI was involved in producing it. Critically, it's cryptographically signed — any modification to the file or the manifest after signing breaks the signature, which makes tampering detectable rather than silently possible, unlike ordinary EXIF metadata.

Who's actually using it in 2026

Adoption has moved from optional to default very quickly. As of this year, Google's Gemini and Imagen image generators embed provenance tags by default, Adobe Firefly signs every generated image with its own credentials, OpenAI's image generation includes full C2PA manifests, and Midjourney has adopted C2PA across its image outputs — a sharp change from a year earlier, when whether AI-assisted work got flagged depended largely on which platform or generator was used.

Hardware is joining in too — some phones now sign photos at the camera level, giving a verifiable chain from capture to publication, not just for AI content but for real photos as proof of authenticity.

How to check it yourself

1. Look for a small **"Cr"** icon on images from supporting platforms (this is the visual indicator for Content Credentials).
 2. Use a free verifier tool — Adobe and the C2PA project both host one — and upload the **original file**, not a screenshot.
 3. The verifier will show you the manifest: creation tool, timestamp, and any recorded edits.
-

The limitations — read this part carefully

This is the step where people most often draw the wrong conclusion, so it's worth being explicit:

- **A missing credential proves nothing.** The vast majority of content — including most authentic photographs and plenty of legitimate AI-generated content from unwatermarked tools — carries no C2PA manifest at all. A bad actor using a non-participating tool produces content that looks, to C2PA, identical to ordinary unsigned content.
 - **Platforms routinely strip it.** One of the largest practical gaps in 2026 is C2PA verification failing on legitimate camera-signed images after platforms re-encode uploads and strip the manifest — which can cause real, authentic photos to be wrongly flagged as AI-generated. If you're checking a social media repost or a downloaded/re-compressed copy, don't be surprised if the credential is simply gone, regardless of the file's true origin.
 - **Presence of a credential isn't a complete guarantee either.** A valid credential doesn't necessarily mean the underlying event happened exactly as implied — it verifies the record, not the full truth behind it.
-

How to check it yourself

1. Look for a small **"Cr"** icon on images from supporting platforms (this is the visual indicator for Content Credentials).
2. Use a free verifier tool and upload the original file — not a screenshot:
 - contentcredentials.org/verify — the official tool from the Content Authenticity Initiative; gives a clean, consumer-friendly summary and validates against the official C2PA Trust List.
 - c2paviewer.com — an independent viewer that shows the raw manifest data (JSON, full assertion breakdown) if you want more technical detail than the official tool surfaces.

- **Digimarc's Content Credentials browser extension (Chrome, Edge, Brave) — checks images automatically as you browse and adds a clickable "Cr" icon when a manifest is present, so you don't have to manually upload each file.**

3. The verifier will show you the manifest: creator/signer, tool used, timestamp, AI-generation indicator, and any recorded edits.

All three run verification client-side — your file isn't uploaded to a remote server, which matters if you're checking something sensitive.

Bottom line for Step 1: if you find a valid, intact Content Credential, that's a strong, verifiable signal — treat it as close to decisive. If you don't find one, that tells you nothing either way, and you move on to Step 2.

Step 2: Scan for an Invisible Watermark

Content Credentials aren't the only layer of built-in provenance. Alongside signed metadata, several AI labs now embed a second, quieter signal directly into the pixels or frames themselves: an invisible statistical watermark. The most widely deployed example is Google DeepMind's **SynthID**.

How watermarking differs from C2PA

C2PA metadata lives in the file's manifest — it's rich (creator, tool, timestamp, edit history) but fragile, because re-encoding, screenshotting, or stripping metadata on upload destroys it entirely. A watermark like SynthID works differently: it's embedded as a subtle statistical pattern in the actual pixel or audio data, so it can survive some of the compression and cropping that would kill a manifest. The two are meant to work together, not compete — one gives you rich context when intact, the other gives you a more durable but narrower confirmation.

Where this stands in 2026

Adoption has accelerated sharply in the last few months. In May 2026, OpenAI joined the C2PA steering committee and committed to embedding SynthID alongside the Content Credentials it already attaches to its image outputs. On the same day, at Google I/O 2026, Google announced that C2PA verification and SynthID detection are being built natively into Search and Chrome — meaning a growing share of everyday browsing will surface this signal automatically, without needing a separate tool.

How to check for it

- If you're using Chrome once the native rollout lands, watermark detection should surface automatically for supported content.
 - Google offers a dedicated **SynthID detector** (via Google DeepMind / Gemini tools) for checking images and audio generated by Google's own models.
 - For now, this check is mostly useful for content you suspect came from Google's Imagen/Gemini or increasingly OpenAI's tools — it isn't a universal scanner for any watermark from any company.
-

The limitations — and this one matters a lot

- **Not every generator watermarks its output.** Midjourney, for instance, doesn't embed SynthID or any other documented watermark — so a clean scan against a Midjourney image will always come back empty, regardless of how the image was made.

- **A missing watermark is not evidence of anything**, for the same reason a missing C2PA credential isn't: absence just means "this tool doesn't participate" or "the signal didn't survive editing," not "this is real."
 - Watermarking works best as **confirmation**, not as a search — you're checking "does this match a known AI signature," not "is this AI at all."
-

"How to check for it" section for Step 2:

How to check for it

- [Google's SynthID Detector](#) — the official verification portal from Google DeepMind. Upload an image, video, or audio file and it scans for a SynthID watermark, highlighting which specific regions or segments triggered a match.
- Gemini app — simpler for casual use: upload the file directly in a Gemini chat and ask whether it was created or altered by Google AI. Gemini checks for both SynthID and C2PA Content Credentials in one pass and gives you a plain-language answer.
- Chrome, Search, and Circle to Search — as of Google I/O 2026, this is now built into the browser itself. In Chrome on desktop you can right-click any image and check if it's AI-generated; on Android, Circle to Search flags AI images automatically. No separate upload needed.
- OpenAI's public verification tool — OpenAI is rolling out its own checker (announced May 2026) to confirm whether an image came from ChatGPT, DALL·E, or the OpenAI API, checking both C2PA metadata and the SynthID watermark OpenAI now embeds jointly with Google.

One important nuance: SynthID detection is closed to Google/OpenAI content only. Google's own tool can't verify Midjourney, Stable Diffusion, or other generators that don't use SynthID — for those, you're relying entirely on Step 3's pixel-analysis detectors instead, since there's no watermark to check for in the first place.

Bottom line for Step 2: if a watermark check comes back positive, treat it as strong confirmation, similar to a valid C2PA credential. If it comes back empty or the tool doesn't cover the platform in question, that's inconclusive — move to Step 3, where you scan the actual pixel/frame content directly rather than relying on anything the generator chose to embed.

Step 3: Run It Through an AI Detection Tool

When provenance metadata is missing and no watermark is found, this is where most people start — a dedicated detector that analyzes the actual pixel or frame content, no metadata required. These tools train classifiers on the statistical fingerprints that generative models leave behind, and unlike Steps 1–2, they don't need the AI tool's cooperation to work.

How these tools actually work

The core method is a trained classifier that scores an upload on a probability scale, and the strongest workflows layer a free first-pass detector, an explainability tool for borderline cases, and an API-based tool for higher volume. Some go further: tools like [Illuminarty](#) generate a heatmap showing exactly which regions of an image triggered detection, which is especially useful when only part of an image was AI-generated — for example, a real photo with an AI-edited background.

Tools worth knowing, by use case

For quick, free, no-signup checks:

- [AI or Not](#) — a free option offering unlimited checks with no signup, good for a fast first-pass read.
- [Is It AI?](#) — offers a limited number of free scans per month for casual single-image checks.
- [WasItAI](#) — no-signup, fast image and audio checks.

For images with detailed forensic breakdown:

- [Illuminarty](#) — provides a probability score plus a heatmap of which regions triggered detection, useful for partially-AI images; it's also notably good at identifying which specific generator or Midjourney version was likely used.
- [FotoForensics](#) — a free forensic tool often used for evidence-grade documentation rather than a simple AI/not-AI verdict.

For video and deepfakes specifically:

- [TrueMedia.org](#) — a free tool aimed at journalists, combining over 10 different detection systems and covering video, image, and audio with roughly 90% accuracy.
- [DuckDuckGoose](#) — processes video in about a second and provides an activation map highlighting suspicious regions, with API access for embedding into video-call workflows.
- [Reality Defender](#), [Sensity](#) — professional-grade deepfake detectors covering audio, image, and video, generally aimed at enterprise/security teams rather than casual use.

For high-volume or developer/API use:

- [Sightengine](#) — works even when metadata has been stripped, which is common for social-platform uploads, and offers a free tier plus paid API plans for teams that need to check images at scale.
 - [Hive Moderation](#) — provides a detailed anomaly breakdown highlighting specific issues like unnatural lighting or texture patterns, alongside identification of which generation model was likely used; priced per-image, which scales well for high-volume users.
-

The accuracy reality check — don't skip this

Numbers here vary a lot by source, generator, and how the test image was processed, so treat any single claimed accuracy figure skeptically. Independent testing gives a useful range: Hive Moderation scored around 94% aggregate accuracy across Midjourney, DALL-E 3, and Stable Diffusion in one 2026 benchmark, while Sightengine hit 98% on some generators but dropped to 75% on others — no tool covers every generator equally well.

More importantly, real-world conditions hurt every tool: an image that scores 90% AI-probability on upload can drop to 40% after a single round of social-media compression. And detectors can be fooled the other way too — heavily processed real photos (HDR composites, aggressive color grading, film grain, anime-style art) routinely trigger false positives, flagging genuine human-made content as AI.

Best practice: don't rely on one tool

The most consistent advice across every source is the same: run more than one detector and treat agreement across tools as the real signal — if results disagree, escalate to a more forensic check rather than trusting either score alone. A practical version of this:

1. Run **one free tool** (AI or Not, WasItAI) for a fast first read.
2. If it's ambiguous or high-stakes, run a **second, different tool** (Illuminarty or Hive) for a cross-check.
3. If the two disagree, or the image looks partially manipulated, use a **heatmap/forensic tool** to see specifically what triggered the flag.
4. Treat the combined result as a **probability, not a verdict** — and move to Step 4 regardless, since visual inspection catches things statistical detectors miss.

Bottom line for Step 3: these tools are the right move when Steps 1 and 2 come back empty, but no score above should be your final answer on its own — especially for anything high-stakes like a legal, journalistic, or reputational decision, where a false accusation carries real cost.

Step 4: Visual and Audio Inspection

This is the manual layer — no software required, just knowing what to actually look for. It's worth doing even after running a detector tool, since automated tools and human eyes catch different things. But this section comes with an important caveat up front: **some of the "classic" advice about spotting deepfakes is now outdated**, and using stale tells can give you false confidence.

The myth to drop: blinking

For years, "watch the blinking rate" was the go-to tip — early deepfakes blinked too rarely because training data was biased toward open-eyed photos. That tell mostly stopped working years ago; modern models blink fine, so relying on it now is more likely to give you false reassurance than to catch anything. If you've seen this advice recently, treat it as outdated.

What still holds up in 2026

Face boundaries and edges. This is one of the more durable tells. Deepfakes typically blend a synthetic face onto existing footage, and that blending can leave artifacts exactly at the seams — blurry or flickering edges where the face meets the hairline, jaw, ears, or neck, sometimes with a subtle "halo" effect. Watch closely when a hand or object passes in front of the face — a common modern tell is the face briefly distorting or "shimmering" as it's occluded and then reconstructed.

Lighting consistency. Check whether shadows on the face match the light source elsewhere in the frame. Synthesized faces are often lit slightly differently from their environment, since the face and background weren't rendered by the same physical process.

Lip-sync alignment. Still one of the most revealing checks. Real speech involves complex, tightly-timed coordination between lips, teeth, tongue, and jaw. Pay particular attention to consonants that require a full mouth closure — "b," "p," and "m" — and to "f" and "v," where the lower lip should contact the upper teeth. If the mouth shape doesn't match those sounds, or lags slightly behind the audio, that's a real signal. Slowing down playback makes these mismatches much easier to catch.

The mute-and-listen trick. A genuinely useful technique: mute the video and watch the lips alone, then close your eyes and listen to the audio alone. Ask whether each one feels natural in isolation. Then play them together — a real recording feels like one unified performance, not two layered tracks stitched together.

Background and physics. Warped or inconsistent backgrounds, text that garbles or shifts between frames, and objects that move with impossible physics or pass through each other are still common weak points, especially in fully AI-generated (not just face-swapped) video.

Audio-specific cues

Voice cloning has become one of the highest-volume categories of AI-driven fraud, so audio deserves its own checklist:

- **Missing natural artifacts.** Real speech includes breath intakes between sentences, lip smacks, and tongue clicks — many synthetic voices omit these entirely, or place them at statistically odd moments.
 - **Mismatched acoustics.** A cloned voice built from one source (say, a podcast clip) can retain that source's reverberation profile even when it's supposedly speaking through a different channel, like a phone call.
 - **Pitch and prosody glitches.** Subtle pitch artifacts at word boundaries, especially on uncommon words or names, and a general flatness or lack of real-time emotional adaptation — AI voices still struggle to genuinely react to interruption or surprise.
 - **Tone-content mismatch.** Check whether the emotional tone of the voice actually matches the facial expression and the content of what's being said.
-

If it's a live video call, not a recording

This is a distinct scenario worth calling out on its own, since real-time face-swap and voice-clone tools are now a documented corporate fraud vector. A few behavioral tests exploit the computational limits of live rendering:

- Ask the person to **turn their head slowly to a full profile angle** — most live face-swap tools blur, warp, or lose tracking at extreme angles.
 - Ask an **unscripted, unexpected question** and watch for a delay or an oddly generic response — real-time generation struggles with genuine spontaneity.
 - For anything involving money or credentials, **hang up and call back on a number you already have on file**, rather than trusting the call itself — this is standard advice from fraud-prevention bodies like the FTC, and it defeats voice cloning entirely since it moves verification to a different channel.
-

Optional forensic tools for this layer

- **DeepFake-o-Meter** — an open academic platform that runs a video through multiple detection algorithms at once, useful if you want a more rigorous second opinion than visual inspection alone.
- **FotoForensics** — uses Error Level Analysis (ELA) to highlight regions of an image that were edited or composited differently than the rest, useful for spotting partial manipulation rather than a fully synthetic image.

The honest limitation of this whole step

It's worth being direct about this: on a well-produced modern fake, human detection accuracy is close to a coin flip. Manual inspection catches low-effort or older-generation fakes reliably, but it's not a strong defense against the current frontier of generators on its own. Treat everything in this section as **one more filter in the stack**, not a verdict — which is exactly why Step 5 (source verification) and Step 6 (weighing signals together) still matter even after a careful visual pass.

Step 5: Verify the Source

Even a technically perfect deepfake has to come from somewhere — and tracing where a piece of content actually originated, who posted it first, and whether it's been corroborated is often what settles a case faster than any pixel-level analysis. This is the step investigative journalists lean on most heavily, and it works even against AI generators the detectors in Step 3 haven't been trained on yet.

Reverse image and video search

Standard reverse search engines work on still images, not video files directly — so for video, the workflow is to pull a clear frame and search that.

- **Google Lens** — the fastest everyday option. In Chrome, right-click any image or video frame and select "Search Image with Google Lens" — it analyzes objects, faces, and text overlays together and often surfaces the original YouTube or TikTok source in seconds. On mobile, Lens now handles short video capture directly.
 - **TinEye** — best for confirming which source published something first, since its "sort by oldest" feature does chronological source tracing rather than just similarity matching — useful evidence when you need to prove something is a repost.
 - **Yandex Images** — often outperforms Google for facial recognition and non-Western content, worth trying as a second pass when Google comes up empty.
 - **Bing Visual Search** — has a crop tool that lets you isolate a specific object, logo, or location within a frame rather than searching the whole image.
-

Professional-grade tools (for higher-stakes verification)

- **InVID-WeVerify** — a free browser extension built for journalists and fact-checkers. It automatically extracts multiple keyframes from a video URL and runs them across Google, Bing, Yandex, TinEye, and Baidu simultaneously, plus offers a magnifier for pixel-level zoom and an experimental deepfake-detection tab. Some advanced features require registering as a journalist or researcher.
 - **Berify** — the one tool that accepts a full video upload directly (rather than requiring manual screenshots) and searches Google, Bing, Yandex, and TinEye at once; aimed at creators and brands monitoring for unauthorized reposts, but works just as well for verification.
 - **Amnesty International's YouTube DataViewer** — extracts the exact upload timestamp and thumbnail history for any YouTube video, useful for confirming when a clip actually first appeared.
-

What to actually check once you find the source

Finding an earlier copy is only the start — the real question is whether the content has been altered or recontextualized along the way. Fact-checkers generally sort issues into two categories:

- **Missing or misleading context** — the footage itself is real and unaltered, but it's being presented with incorrect framing, or a brief clip has been pulled from a much longer video in a way that creates a false narrative.
 - **Deceptive editing** — the video has actually been cut or rearranged: large portions edited out to skew the story, or different clips spliced together to fabricate an event that never happened as shown.
-

Practical checks:

- Who posted it **first**, and does that account have a track record?
 - Have **reputable outlets** picked it up and corroborated it independently?
 - Does the video's content **align with other known facts** — is the person saying something that contradicts verified past statements?
-

The honest limitation here

Reverse search fails on brand-new content that hasn't been indexed yet, on heavily edited or cropped clips, and on anything posted to private or paywalled platforms outside public search engines' reach. It's a powerful step, but like every other one in this guide, it's a signal to weigh — not a final verdict on its own.

Step 6: Weigh All the Signals Together

This is where the process actually resolves. None of the previous five steps was designed to give you a final verdict on its own — each one is a filter, and confidence comes from how many of them agree, not from any single one crossing a threshold.

A simple framework for combining signals

Think of each step as contributing evidence in one of three directions: **confirms real**, **confirms AI**, or **inconclusive**. Most steps will land on inconclusive most of the time — that's normal, not a failure of the process.

Step	Strong "AI" signal	Strong "real" signal	Inconclusive (move on)
1. C2PA credential	Valid manifest tags it as AI-generated	Valid manifest shows camera capture, intact signature	No manifest found at all
2. Watermark (SynthID)	Watermark detected	— (absence isn't evidence)	Tool doesn't cover this platform/generator
3. Detection tool	Multiple tools agree, high probability score	Multiple tools agree, low probability score	Tools disagree, or score is mid-range
4. Visual/audio inspection	Clear edge artifacts, lip-sync mismatch, lighting inconsistency	Nothing found after careful review	Ambiguous, hard to tell either way
5. Source verification	No credible origin, recently surfaced with no corroboration	Established source, corroborated by reputable outlets	Can't trace origin either way

A confident call requires agreement across at least two or three independent steps, ideally ones that work on different principles (e.g., a metadata signal plus a pixel-analysis signal plus a source check). A single strong signal from just one step — especially a detector score, which can swing wildly with compression — isn't enough to act on for anything that matters.

How much rigor you actually need depends on the stakes

Not every situation needs the full six-step process. Match the effort to what's riding on the answer:

- Casual scrolling / deciding whether to share something** — Steps 1–2 (quick provenance/watermark check via Chrome or Gemini) plus a glance at the visual cues in Step 4 is usually enough. If it looks off and you can't verify it, just don't share it.

- **Content you're about to publish or report on** — Run the full sequence: provenance, watermark, at least two detection tools, careful visual/audio inspection, and source verification with corroboration from a second outlet.
 - **High-stakes decisions** (financial transfers, hiring decisions, legal or security matters, an urgent request from someone claiming to be an executive or family member) — Treat any unsolicited video or call as a risk signal, not proof, regardless of how convincing it looks. Verify through an out-of-band channel — call the person back on a number you already have — rather than trusting the media itself at all.
-

A word on false accusations

It's worth stating plainly: wrongly calling a real photo or video "fake" carries real cost too — to someone's reputation, to a legitimate news story, to a legal case. Detectors have documented false-positive rates on heavily processed but genuine content (HDR photos, aggressive color grading, older film grain), and C2PA metadata frequently gets stripped from real camera footage by ordinary platform re-encoding. The same discipline that stops you from believing an obvious fake should also stop you from confidently declaring something fake on the strength of one ambiguous signal.

The bottom line

Detection and generation are in a permanent arms race, and detection will structurally always lag a step behind — new generator architectures routinely slip past classifiers trained on older ones, and provenance standards only work when the tool that made the content chose to participate. That's not a reason to give up on verification; it's the reason no single tool or check should ever be treated as the final word.

The six-step process in this guide isn't about achieving certainty — it's about building a defensible, repeatable habit: check provenance, check for a watermark, run a detector or two, look and listen carefully, trace the source, and weigh it all together before you trust, publish, or act on what you're seeing. In a media environment where "seeing is believing" no longer holds, that habit is the actual skill worth having.
