THE DEEPFAKE DETECTION HANDBOOK

A Practical Guide to Spotting AI-Generated Text, Images, Video, and Voice

For journalists, fact-checkers, professors, researchers, and citizens who care about the truth.

By Umut İlhan
Valmera AI Consulting & Training
April 2026 Edition

Contents

Introduction: Why This Book Exists

In 2026, anyone with a smartphone can generate a convincing fake photo in fifteen seconds, clone a voice from three seconds of audio, write a fake press release in two minutes, or fabricate a video of an event that never happened. The barriers to creating synthetic media have collapsed. The barriers to detecting it have not.

This handbook exists because the people who most need detection skills — journalists working under deadline, professors grading hundreds of essays, fact-checkers facing viral content, and ordinary citizens who just want to know what's real — cannot rely on automated detection tools. Those tools fail constantly. They flag real human writing as AI. They miss obvious deepfakes. They give false confidence to people who shouldn't trust them.

The good news is that AI-generated content leaves traces. Not perfect ones, not always obvious ones, but consistent ones. A trained human eye can spot these traces faster and more reliably than any commercial detector. This book teaches you what to look for, where to look, and how to weigh what you find.

You will not become infallible. No one is. But you will become competent — competent enough to make better decisions about what to publish, what to teach, what to share, and what to trust.

Who This Book Is For

- Journalists who need to verify content under deadline pressure
- Fact-checkers working with viral social media content
- Professors and teachers evaluating student submissions
- Researchers studying disinformation
- Public relations professionals defending their organisations
- Citizens who want to think critically about what they see online

What This Book Is Not

This is not a software manual. The tools mentioned here will change. Some will disappear. New ones will replace them. What will not change is the methodology: the questions to ask, the signals to find, the discipline to layer multiple checks.

This is not a guarantee. Detection is probabilistic. Anyone who tells you they can spot 100% of AI content is selling something. The best you can achieve is high-confidence judgment based on multiple converging signals.

This is not about catching cheaters or accusing people. The same techniques that detect malicious deepfakes also detect honest mistakes, professional polish, and stylistic accidents. Use this knowledge with humility.

How to Use This Handbook

Each detection chapter follows the same structure, so you can move between modalities without relearning the format:

1. Threat overview — what kind of fake exists in this medium and what damage it can do
2. The signal categories — what to look for, organised by type
3. Step-by-step inspection workflow — exactly how to examine the content
4. Scoring framework — how to combine signals into a confidence score
5. Worked examples — real cases with full reasoning
6. The Liar's Dividend warning — what this method cannot detect

Read it linearly the first time. After that, treat it as a reference. When you receive a suspicious image, jump to Chapter 3. When you encounter a strange essay, jump to Chapter 2. The chapters are self-contained.

Practice matters more than reading. After each chapter, find five real examples — three you suspect are AI, two you know are human — and run the methodology. Compare your judgments with the ground truth. Calibrate your eye over time.

Chapter 1 — The Mindset: Zero-Trust Verification

Trust nothing by default. Earn confidence through evidence.

Before any technique, you need a mindset. The Zero-Trust principle is simple: treat every piece of media as potentially synthetic until you have positive evidence otherwise. This is not paranoia. It is professional discipline.

The opposite mindset — assuming content is real until proven fake — was acceptable when fakery required skill, time, and money. In 2026, fakery requires none of those. The cost of generating a convincing fake image dropped from hundreds of dollars and a Photoshop expert to zero dollars and ten seconds. Your default has to change to match.

The Three Threats

There are three different ways media can deceive you, and each requires a different response.

Threat One: Mislabelling

Real content presented as something it isn't. A real photo from a 2018 protest passed off as today's. A real video of one event captioned as another. This is the oldest form of disinformation. Detection requires reverse search and source verification, not pixel analysis.

Threat Two: Editing

Real content modified to change its meaning. A photo with an object added, a video with audio replaced, a quote with words changed. Detection requires forensic analysis of the file itself.

Threat Three: Full Synthesis

Content that never existed in the real world. AI-generated photos of events that never happened, voices saying things never said, videos of people doing things they never did. Detection requires understanding how AI generation works and where it fails.

The Liar's Dividend

Every detection capability creates an equal and opposite risk: real content can be dismissed as fake. When deepfakes become possible, real videos of real events become deniable. A politician caught on tape can simply claim the tape is a deepfake. A protester documented committing violence can deny the documentation.

This is the Liar's Dividend, and it is at least as dangerous as deepfakes themselves. Your job is not just to catch fakes — it is to confidently authenticate real content as well. A score of 'suspicious' on a real video is just as harmful as a score of 'authentic' on a fake one.

Throughout this handbook, you will see warnings about this dividend. Take them seriously. The instinct to assume everything is fake is just as wrong as the instinct to assume everything is real.

Chapter 2 — Detecting AI-Generated TEXT

The Threat

AI text is the most common form of synthetic content. It powers fake press releases, fake academic papers, fake reviews, fake quotes attributed to real people, fake comments shaping public opinion, and ghost-written content presented as original. By late 2024, an estimated 40% of biomedical abstracts in some scientific fields were AI-assisted. By 2026, the question is no longer 'is this AI?' but 'how much AI is in this?'

Worse, automated text detectors do not work. Grammarly, ZeroGPT, GPTZero, Originality.ai — all tested by independent researchers, all found unreliable. They fail in two opposite ways: they flag real human writing as AI (especially work by professional writers and non-native speakers), and they miss AI text that has been even slightly edited. The 'undetectable AI' services that charge $20/month are selling protection against a threat that algorithms cannot reliably catch in the first place.

So you have to use your eyes. The good news: trained eyes are far more reliable than any tool.

The Eight Signal Categories

Signal 1 — Vocabulary Tells

AI overuses specific words at frequencies impossible for natural human writing. The watchlist:

- Verbs: delve, foster, showcase, harness, unveil, underscore, represents
- Adjectives: intricate, pivotal, vibrant, meticulous, profound
- Nouns: tapestry, landscape (used abstractly), realm, testament, symphony
- Phrases: 'stand as a testament', 'rich tapestry of', 'in the realm of', 'moving beyond'

One word from this list means nothing. Real humans use these words occasionally. Three or more in a 200-word passage is a strong signal. Seven or more in an article is near-certain AI.

Signal 2 — Triplets (The Master Key)

This is the strongest single AI tell. AI is obsessed with listing things in threes. Look for three adjectives stacked ('fast, reliable, and efficient'), three short sentences in a row, three bullet points in a row, three -ing verbs hanging off the end of a sentence ('reshaping industries, transforming societies, redefining work').

Real writers occasionally use the rule of three for rhetorical effect. AI uses it constantly because it learned from rhetorical training data and never forgot. If you see two or more triplet structures in a single paragraph, you are almost certainly looking at AI output.

Signal 3 — Reversal Structures

The 'not just X, it's Y' formula. AI loves it. Humans rarely use it. Examples: 'It is not just a product, it is a paradigm shift.' 'She doesn't just sit, she poses.' 'This isn't simply a tool, it is a revolution.'

Real people don't usually feel the need to tell you what something IS NOT before telling you what it IS. One reversal is tolerable. Two or more in a short text is a strong AI signal.

Signal 4 — Modular Sentences

Real human writing has flow. Each sentence builds on the previous one through connectors: 'because', 'while', 'however', 'so', 'which'. AI writing is a list of statements pretending to be a paragraph. Each sentence stands alone.

The test: take any paragraph. Mentally swap sentence 1 with sentence 3. Does the meaning still work? If yes, the writing is modular, which is a strong AI signal. Also count connector words: fewer than two per 100 words is suspicious.

Signal 5 — Missing Hedging

Real humans signal uncertainty. They say 'I think', 'maybe', 'I suspect', 'as far as I know', 'I could be wrong but'. Linguists call this 'epistemic hedging'. It is human. AI rarely does it because it is tuned to sound confident — users reward confident answers, so AI companies trained their models to be confident.

A 100-word text with zero hedging expressions is a signal. A 500-word essay with no 'I think' or 'maybe' or 'in my experience' is a strong signal. Note: this signal is weaker for formal academic writing, where hedging is often discouraged.

Signal 6 — Sophistry

This is the most important and most sinister signal. Sophistry is writing that sounds well-reasoned and confident but doesn't actually say much. Big words disguising empty content.

Example: 'The integration of artificial intelligence into modern workflows represents a transformative paradigm shift, fundamentally reshaping the very fabric of how organisations approach knowledge creation.' Now translate to plain English: 'AI is changing how companies work.' Five words capture the entire thought.

The test: pick any paragraph. Ask 'can I summarise this in six words?' If yes and the original is fifty or more words, you are looking at sophistry. AI produces sophistry constantly because it was trained on corporate marketing copy, motivational LinkedIn posts, and consultant slide decks — all of which run on sophistry.

Signal 7 — Structural Defaults

AI has a recognisable format pattern: 800 to 1000 total words, 4 to 6 paragraphs, 150 to 250 words per paragraph, 3 to 5 sentences per paragraph. Other structural tells include heavy em-dash usage (—), echoing sentence patterns across paragraphs, and dramatic capitalisation of ordinary nouns ('The Witness', 'Corporate Memory').

None of these alone is conclusive. Combined, they form a recognisable shape. A 'press release' that fits this exact mould should be examined more carefully than one that doesn't.

Signal 8 — Tonal Markers

Overpolished language. No personal anecdotes. Robotic transitions. Authoritative confidence with no personal stake. No comparatives or superlatives — AI prefers 'noteworthy' and 'significant' over 'best' and 'worst'. No 'I' or 'we'. No emotional warmth even on emotional topics.

These are subjective, but trained eyes recognise them. Read enough AI text and the pattern becomes obvious.

Step-by-Step Inspection Workflow

7. Read the text once at normal speed. Form a gut impression.
8. Scan for the eight vocabulary watchwords. Count occurrences.
9. Highlight every triplet structure you can find.
10. Search for 'not just X, it's Y' patterns.
11. Pick a paragraph. Mentally swap sentences. Test for modularity.
12. Count connector words per 100 words of text.
13. Search for hedging expressions — 'I think', 'maybe', 'perhaps'.
14. Pick a paragraph and test for sophistry — six-word summary.
15. Note structural metrics: word count, paragraph count, em-dash count.
16. Sum the signals. Apply the scoring table below.

Scoring Framework

Count the categories (1 to 8) where you found a flag. Apply this table:

- 0-1 flags → 1 to 30 score range → Likely human
- 2 flags → 30 to 45 → Uncertain
- 3 flags → 45 to 65 → Suspicious
- 4 flags → 65 to 80 → Likely AI
- 5 flags → 80 to 90 → Very likely AI
- 6+ flags → 90 to 99 → Almost certainly AI

Adjustments: subtract 10 to 15 points for academic, technical, or legal writing. Subtract 10 to 20 for translated text. Subtract 10 to 15 if the writer is clearly non-native. Never assign 0. Never assign 100.

Worked Example

Paragraph to analyse:

> '*The rise of artificial intelligence represents one of the most profound transformations in human history, reshaping industries, societies, and even our understanding of what it means to be intelligent. In the past decade, it*

has evolved from a niche academic pursuit into a central pillar of the global economy and culture. The combination of big data, computational power, and advances in machine learning algorithms has enabled it to move beyond theoretical models and into real-world applications.'

Signal analysis:

- Vocabulary: 'represents', 'profound', 'moving beyond' — three watchlist hits — FLAG
- Triplets: 'reshaping industries, societies, and... understanding' (three) and 'big data, computational power, and machine learning' (three) — FLAG
- Reversal structures: none found — PASS
- Modular sentences: each sentence stands alone, sentence 1 and sentence 3 could swap — FLAG
- Hedging: zero — FLAG
- Sophistry: entire paragraph summarises as 'AI grew' — FLAG
- Structural: not enough data from one paragraph
- Tonal: bland confidence, no personal voice — FLAG

Six flags. Score: approximately 88 / 99. Verdict: almost certainly AI-generated.

Chapter 3 — Detecting AI-Generated IMAGES

The Threat

AI image generation became consumer-grade in 2023 and photorealistic in 2024. By 2026, tools like Google Nano Banana, Midjourney, and FLUX produce images indistinguishable from photographs at first glance. They generate fake protest scenes, fake disasters, fake intimate images, fake historical photos, and fake portraits of real people.

Unlike text, images carry pixel-level information that can be analysed forensically. The good news: AI generators leave consistent visual signatures. The bad news: those signatures change as the technology improves. The fingers and hands that betrayed AI in 2023 are mostly fixed in 2026. New signatures take their place.

This chapter teaches you the eight inspection categories that work in 2026. The categories will outlast specific signatures — what changes is which category catches the most fakes in any given year.

The Eight Inspection Categories

Category 1 — Hands, Fingers, and Anatomy

Still the most reliable category. Count fingers when they are visible. Real hands have five fingers. AI hands often have four or six, fingers that merge into each other, fingers that pass through objects they should grip, or hands holding objects in physically impossible ways.

Beyond hands: count limbs. Look for extra arms in crowds, legs in unnatural positions, joints bending the wrong way. Check teeth — AI teeth are often too uniform, too many, or identical in shape (real teeth vary). Check eyes for asymmetric pupils, mismatched iris colours, missing reflections, or a 'dead' stare. Check ears — AI often produces mismatched ears on the same person.

Category 2 — Lighting and Shadows

Real photographs obey physics. AI images sometimes don't. Check that all shadows in the image point away from a single light source. If the sun is on the left, every shadow should fall to the right. Check that shadow softness matches the light source — sharp shadows mean direct sun, soft shadows mean overcast or diffused light. Mixed shadow types in the same image are a red flag.

Check reflections. Eyes should reflect the dominant light source. Glass, water, and metal should show consistent reflections. AI often gets eye reflections wrong — they appear in the wrong eye, or are missing entirely.

Category 3 — Edges, Boundaries, and Hair

Where AI generation falls apart. Look at the edge between hair and background. Real photographs show crisp individual hair strands or a soft photographic blur

from depth of field. AI shows a melted, gradient-blurred 'halo' that is neither sharp nor naturally soft.

Look at jawlines and ears against backgrounds. Look at glasses frames — they often warp, become asymmetric, or appear to float. Look at clothing edges — collars, cuffs, and hems often smear into surrounding skin or fabric.

Category 4 — Text and Symbols

AI's biggest weakness in 2026 is still text. Any visible writing in an image is a verification opportunity. Read every sign, every label, every button, every license plate, every book spine. AI text often looks correct at first glance but reveals as gibberish on closer inspection. Letters distort, words spell nothing, numbers misalign, logos warp.

Watches and clocks are particularly diagnostic. AI struggles to render coherent clock faces. If an image contains a clock, examine the numbers and hands carefully.

Category 5 — Skin Texture and Faces

AI faces have a signature look. Skin is too smooth — no pores, no asymmetry, no individual variation. Real faces are slightly asymmetric; AI faces are often perfectly mirrored. The hairline transition is often too clean, like a 'beauty filter' was applied to the face only.

In images containing crowds, look for repeated faces. AI often duplicates the same face across multiple background characters with minor variations.

Category 6 — Background Anomalies

AI focuses computational effort on the main subject. Backgrounds get less attention. Look for repeating patterns where there shouldn't be any. Look for impossible building geometry — windows that don't align, roofs that don't connect, walls at impossible angles. Look for ghostly figures — partial people or half-formed bodies in the distance.

Category 7 — Context, Physics, and Logic

Does the scene make logical sense? Wet ground but dry clothes. Snow on the roof but green leaves on the trees. Modern cars in a 1950s street scene. Weather that doesn't match the season. Cultural mismatches — Western signs in an Asian city, or vice versa.

This category rewards general knowledge. The more you know about the claimed location, era, and context, the more anomalies you will catch.

Category 8 — Provenance Markers

Some AI tools leave visible signatures. Google Gemini and Nano Banana embed a small ✦ sparkle icon in the bottom-right corner. DALL-E embeds a coloured square in the bottom-right. Midjourney has a distinctive aesthetic and usually exports at specific aspect ratios.

Filename context also matters. If the file is named 'Generated_Image_2026...' or similar, that is a strong hint, though filenames can be changed. Image metadata

(EXIF) can reveal the originating software — but absent EXIF doesn't prove AI; many real images have EXIF stripped during social media upload.

Step-by-Step Inspection Workflow

17. View the image at full resolution. Never judge from thumbnails.
18. Scan the bottom-right corner for visible AI watermarks (✦, square).
19. Find every hand in the image. Count fingers.
20. Identify the light source. Trace shadows back to it. Check consistency.
21. Examine hair-to-background transitions for melting.
22. Read every visible piece of text. Look for distortion or gibberish.
23. Examine faces for symmetry and skin texture.
24. Scan the background for repeated patterns or impossible geometry.
25. Apply common sense — does the scene logically work?
26. Check filename and EXIF if available.
27. Sum the flags and apply the scoring framework.

Scoring Framework

- 0-1 flags → 1 to 30 → Likely authentic
- 2 flags → 30 to 45 → Uncertain
- 3 flags → 45 to 65 → Suspicious
- 4 flags → 65 to 80 → Likely AI
- 5 flags → 80 to 90 → Very likely AI
- 6+ flags → 90 to 99 → Almost certainly AI

Visible AI watermark: add 20 points (cap at 99). Heavy compression on social media: cap confidence at medium. Artistic illustration: standard rules don't apply, warn before scoring.

The Forensic Tool Layer

Manual inspection gives you the score. Forensic tools help you double-check before publication. The free, browser-based tools to know:

- Forensically (29a.ch) — runs ELA, clone detection, noise analysis
- InVID WeVerify — provides keyframe extraction and a 'Deepfake' detection tab
- FotoForensics — quick ELA check
- Google Lens, Yandex, TinEye, Bing Visual Search — for reverse image search

Always run reverse image search first. If the image already exists somewhere on the web before its claimed context, it is mislabelled, not AI — and that is a different problem.

Chapter 4 — Detecting AI-Generated VIDEO

The Threat

Video synthesis crossed the photorealism threshold in 2024 with OpenAI's Sora. By 2026, tools like Google Veo, Kling, and Seedance produce 4K video at 60 frames per second with native synchronised audio. They generate fake speeches by world leaders, fake protest footage, fake disaster scenes, and fake interview clips.

Video is the most dangerous form of synthetic media because it carries the highest social trust. People who learned to distrust photos still trust video. They shouldn't.

The good news: video adds two dimensions images lack — time and audio. Both create new failure modes for AI. Master this chapter and you can spot AI video faster than AI image, even though the underlying technology is more sophisticated.

Four Types of Video Manipulation

Before you analyse, identify which type you are dealing with. Different types require different techniques.

- Full Synthetic — entirely AI-generated, no real footage involved
- Face Swap — real video with one person's face replaced by another's
- Lip Sync — real face and expression with new audio and matching mouth movements
- Object Manipulation — real video with specific elements added, removed, or changed

The Ten Inspection Categories

Category 1 — Temporal Consistency (The Master Key)

Always check this first. AI video models generate frames in groups, and statistical noise creates micro-variations between frames that real cameras never produce. Look for objects that subtly change shape from one frame to the next. Look for edges that flicker or shimmer. Look for backgrounds that morph slowly through the clip. Look for textures that loop or repeat.

Use the period (.) key on YouTube to step through video frame by frame. Pay attention to static objects — buildings, signs, furniture. They should be perfectly stable. Any flicker, any drift, any colour shift on a stationary object is a strong signal.

Category 2 — Faces and Lip Sync

For talking-head video, this is the second most important check. Watch the lips against the audio. Real lip movement has specific shapes for specific sounds — 'p' and 'b' close the lips, 'f' and 'v' touch the lower lip to upper teeth, 'th' shows the tongue. AI often gets these wrong, or gets the timing slightly off.

Look for boundary artifacts at the jawline — the classic face-swap signature. Look for skin-texture mismatches between face and neck. Look at blinking patterns — too regular, too rare, or absent.

Category 3 — Hands, Limbs, and Anatomy in Motion

AI video models still fail at extremities, especially during movement. Watch hands when they enter the frame. Count fingers. Watch limbs during fast movement — they sometimes pass through solid objects, bend wrong, or distort. Watch heads during rotation — AI sometimes rotates them past the natural 180-degree limit before the body follows.

Category 4 — Physics and Motion

Gravity, inertia, and material physics are hard to fake. Watch falling objects — do they accelerate naturally? Watch cloth — does it drape and flow with believable weight? Watch hair — does it move with the head or trail behind unnaturally? Watch water and smoke — do they behave like fluids?

The walking-gait test is particularly diagnostic. Real human walking involves subtle weight shifts, slight hip rotation, natural arm swing. AI walking is often too smooth, too symmetric, or robotically mechanical.

Category 5 — Lighting and Shadows Across Time

Same principles as image inspection, but with the added test of consistency across the clip. As subjects move through the scene, their shadows should move predictably. As objects rotate, their highlights should shift correctly. AI sometimes maintains lighting consistency within a single frame but fails to update it correctly as motion progresses.

Category 6 — Background and Scene Stability

Watch the background, not the subject. Background people in AI video are often partial, ghostly, or anatomically incoherent. Background text morphs frame to frame. Background buildings shift their geometry. Background vehicles have wheels that don't turn correctly.

Category 7 — Audio-Visual Synchronisation

If the video has audio, treat sync as a critical signal. Footsteps should match footfalls. Breaking glass should sound when glass breaks. The room ambience should match the visual environment — outdoor scenes need wind, indoor scenes need echo characteristics. Voice quality should match microphone placement.

Category 8 — Resolution and Quality Patterns

Real cameras produce grain. AI video is often unnaturally clean. Check whether sharpness is consistent across the frame — face sharp but edges of frame soft is a signal. Check frame rate stability — sudden judders suggest frame interpolation or generation switching.

Category 9 — Multi-Shot Coherence

For edited videos with cuts, check continuity. Same person across cuts should have identical clothing, identical wrinkles, identical lighting (within reason). Background details should remain consistent. Time-of-day cues should match. Inter-scene continuity is where models like Seedance break down.

Category 10 — Provenance Markers

Google Veo videos may carry the ✦ sparkle icon or invisible SynthID watermarking. Sora videos often have a recognisable aesthetic. Kling videos sometimes carry visible Chinese branding. Filename context matters.

Step-by-Step Inspection Workflow

28. Watch the video at normal speed once. Form a gut impression.
29. Watch again with audio muted. Focus on visuals only.
30. Watch frame-by-frame using the period key on YouTube.
31. Identify static objects. Confirm they remain stable across all frames.
32. Watch faces and lips against audio carefully.
33. Check hands and limbs during movement.
34. Test physics — is gravity, inertia, and motion natural?
35. Watch background people, signs, and architecture.
36. Listen with audio only — does ambience match visuals?
37. Check for visible watermarks or filename context.
38. Sum the flags and apply the scoring framework.

Scoring Framework

- 0-1 flags → 1 to 30 → Likely authentic
- 2 flags → 30 to 45 → Uncertain
- 3 flags → 45 to 60 → Suspicious
- 4 flags → 60 to 75 → Likely AI
- 5 flags → 75 to 87 → Very likely AI
- 6+ flags → 87 to 99 → Almost certainly AI

Adjustments: temporal inconsistency triggers a +10 boost. Lip-sync failure on a talking head triggers a +15 boost. Visible watermark adds +20. Very short clips (under 3 seconds) cap confidence at medium. Animation, CGI, and game capture do not score on this scale — they are intentionally non-photorealistic.

Chapter 5 — Detecting AI-Generated VOICE

The Threat

In January 2024, voters in New Hampshire received automated phone calls in what sounded clearly like President Joe Biden's voice, telling them not to vote in the primary. The calls were synthetic. They cost less than five dollars to produce and took under twenty minutes. The perpetrator was eventually fined six million dollars and faced twenty-four criminal charges. Before he was caught, an unknown number of voters stayed home.

That same year, criminals used voice cloning to impersonate a CEO and authorise a wire transfer of over two million dollars. The CEO's own finance team could not tell the difference.

Modern voice cloning needs about three seconds of source audio. Three seconds. That means anyone with a public voice — every journalist who has ever appeared on television, every politician who has spoken in public, every CEO who has given an interview — is a potential target.

And here is the disturbing part: a 2025 Berkeley study found that humans correctly identify voice clones only about 50% of the time. Barely better than guessing. Your ears are not enough.

The Eight Inspection Categories

Category 1 — Breathing and Biology (The Master Key)

This is the strongest single tell. Real human speech is biological. It contains tiny sounds that AI often forgets: audible inhales before long sentences, soft exhales at the ends of phrases, mouth clicks between words, lip smacks, occasional swallows, throat clearings, the natural energy decay at the end of syllables.

The test: listen to thirty consecutive seconds of speech. Did you hear any breathing? Any mouth sounds? If the answer is no — if the audio is biologically silent — that is the strongest possible AI signal. Real humans cannot speak for thirty seconds without their bodies making sound.

Category 2 — Pitch and Prosody

Real voices have natural micro-fluctuations in pitch — what audio engineers call 'jitter', usually 5 to 12 random variations per second. AI voices generate pitch through mathematical curves that are statistically too smooth. The difference is impossible to consciously hear, but you can feel it: AI voices sound 'too perfect'.

Listen for pitch variation matching meaning. Real questions rise at the end. Real emphasis falls on key words. Real emotion changes the pitch curve. AI often delivers everything with the same melodic shape regardless of content.

Category 3 — Pauses and Rhythm

Real people hesitate. They restart sentences. They say 'um' and 'uh' and 'you know'. Linguists call these 'disfluencies' — and disfluencies are HUMAN. Their absence is suspicious.

Listen for 60 seconds of speech and count the filler words. Zero fillers in 60 seconds of natural speech is a strong AI signal. Pause length should also vary based on meaning, not be uniform. Pause placement should fall at natural grammatical boundaries, not random points.

Category 4 — Articulation Precision

Listen to consonants. Real speakers slur slightly. They mumble parts of words. They have accents and personal quirks. AI voices often produce every consonant with mechanical precision — every P perfectly plosive, every S perfectly sibilant, every T perfectly crisp.

Real English has 'gonna', 'wanna', 'kinda'. Real Turkish has dropped vowels and casual contractions. AI tends toward textbook pronunciation. Robotic precision is a signal.

Category 5 — Emotional Authenticity

The hardest thing for AI to fake. Real emotion produces specific vocal effects: voice cracks under sadness, tremor under fear, throat tightening under stress, breathing changes with excitement. Real speakers cry, laugh, gasp, sigh, and let their voices break.

If the content is emotional but the delivery is flat, that mismatch is a strong signal. AI voices tend toward neutral delivery even on highly emotional material.

Category 6 — Drift Over Time

For recordings longer than thirty seconds, listen for subtle changes. Does the speaker sound the same at second 5 as at second 45? AI voice generation sometimes drifts: speaking rate slowly speeds up or slows down, emotional engagement fades, accents shift slightly, energy levels vary in unnatural ways.

Real speakers maintain a consistent personal baseline because their physical bodies don't change in a minute. AI parameters can drift across a single generation.

Category 7 — Background and Environment

Real recordings exist in a context. Every microphone picks up some ambient sound — air conditioning, distant traffic, room tone, the hum of electronics. Real speakers move while talking, producing rustles and small body sounds. Real rooms have echo characteristics.

AI voices are often unnaturally clean. A perfectly silent background with a voice in the foreground is a signal. Real recordings have an 'audio floor' — a quiet baseline that exists even in silence.

Category 8 — Technical and Provenance Markers

Spectral analysis can reveal signatures in the 4 to 8 kHz frequency range that AI voices often produce as artifacts. The free tool Sonic Visualiser lets you view a spectrogram and look for repeating patterns in this band.

Listen for sibilance handling — real S sounds have natural variation; AI S sounds are sometimes mechanically smoothed. Listen for clip duration tells — AI voice clips are often suspiciously round numbers (10s, 30s, 60s). The most common voice cloning tool, ElevenLabs, has a recognisable 'polish' that experienced ears learn to detect.

Step-by-Step Inspection Workflow

39. Listen to the full clip once at normal speed.
40. Listen again at 0.75x speed. Slower playback exposes biology mistakes.
41. Count audible breaths in the first 30 seconds.
42. Count filler words ('um', 'uh', 'you know') in 60 seconds.
43. Listen for pitch naturalness — does it sound 'too smooth'?
44. Listen to consonant precision — over-articulated or natural?
45. Listen for emotional authenticity if content is emotional.
46. Listen for drift across the clip.
47. Listen to background — silent or naturally ambient?
48. If possible, view the spectrogram in Sonic Visualiser.
49. Sum the flags and apply the scoring framework.

Scoring Framework

- 0-1 flags → 1 to 30 → Likely authentic
- 2 flags → 30 to 45 → Uncertain
- 3 flags → 45 to 60 → Suspicious
- 4 flags → 60 to 75 → Likely AI
- 5 flags → 75 to 87 → Very likely AI
- 6+ flags → 87 to 99 → Almost certainly AI

Adjustments: no breathing in 30+ seconds adds +15. Zero fillers in 60+ seconds adds +10. Completely silent background adds +10. Drift across the clip adds +10. Phone-quality audio caps confidence at medium. Singing or rapping is not standard speech — apply with caution. Professional narrators (audiobooks) naturally produce polished delivery — subtract 10 to 15.

Chapter 6 — The Universal 5-Step Verification Protocol

All four detection methodologies share a common skeleton. When you face any unknown piece of media, run this five-step protocol regardless of the medium:

Step 1 — Lateral Reading

Always start here. Before you analyse the file itself, ask: does this content already exist somewhere else? Run reverse image search on photos. Search distinctive phrases from text. Search the audio waveform if you have transcription tools. Check whether the content has been reported on by trusted sources.

Most disinformation is not deepfake content. It is real content mislabelled or taken out of context. Lateral reading catches this faster than any forensic analysis.

Step 2 — Provenance

Find the original. When was this first published? Who first published it? What was the original context? Track the chain of custody backward as far as possible. The further you can trace, the more confident your final judgment.

Step 3 — Technical Forensics

Now apply the methodology specific to the medium. For text, run the eight signal categories. For images, examine the eight inspection points. For video, check the ten temporal and spatial markers. For voice, listen across the eight acoustic categories.

Generate a score. Generate reasoning. Document what you found.

Step 4 — Geospatial and Temporal Verification

Does the content match its claimed time and place? For images and video, check shadow angles against sun position calculators (SunCalc.org is free). Check architecture against street view. Check vehicles, signage, language, and cultural details against the claimed location. Check timestamps against weather records and known events.

This step is where Bellingcat-style open-source intelligence meets forensic analysis. It is also where mismatches become unambiguous.

Step 5 — Editorial Decision

Combine everything into a final verdict. Use a three-level system:

- GREEN — verified, multiple independent confirmations, safe to publish
- AMBER — uncertain or insufficient evidence, hold for further investigation
- RED — strong evidence of manipulation, do not publish

When in doubt, choose AMBER. Publishing a false GREEN is worse than publishing nothing. The goal of verification is not speed — it is being right.

Chapter 7 — The Liar's Dividend and Editorial Responsibility

Everything you have learned in this book can be used wrongly. The same techniques that catch fake content can be turned against real content. The same scoring frameworks that protect against deepfakes can be weaponised against the truth. This chapter is about responsibility.

What Detection Cannot Tell You

Detection methods identify patterns. They do not identify truth. A high AI score on a piece of writing does not prove the writer used AI — it proves the writing matches statistical patterns common in AI output. Many real human writers, especially professionals and non-native speakers, naturally produce content that matches those patterns.

Never accuse anyone based solely on a detection score. The score is a starting point for investigation, not a conviction.

Who Gets Wrongly Flagged

Detection methods systematically over-flag certain groups. Be aware of who:

- Non-native English (or Turkish) speakers — formal style mimics AI
- Professional writers and editors — years of training produce consistent style
- Academic researchers — formal academic style is statistically predictable
- Translators — translation reduces stylistic variation
- Students with strong writing skills — structured essays trigger false positives
- Public figures with media training — polished delivery sounds artificial

If your detection score targets someone in one of these groups, apply extra scepticism. Consider whether the score reflects actual AI use or merely the cost of being skilled.

The Two Failure Modes

There are two opposite ways to fail at verification. Both cause damage.

False Positive: Calling Real Content Fake

Damage: dismissing real evidence, denying real victims their voice, granting cover to wrongdoers who claim authentic recordings of their misdeeds are 'just deepfakes'.

False Negative: Calling Fake Content Real

Damage: amplifying disinformation, lending institutional credibility to lies, contributing to public confusion about basic facts.

Most professional discussions focus on false negatives. False positives deserve equal attention. The best verifiers are calibrated for both.

The Three Rules

50. Never broadcast unverified content under time pressure. Speed is the enemy of accuracy. Competitors can publish first; you publish correctly.
51. Never rely on a single tool, single signal, or single check. Layer multiple methods. Require convergent evidence.
52. Always state your confidence level. 'We could not verify' is a legitimate position. So is 'we believe this is authentic but cannot prove it.' Pretending to certainty you don't have is the path to professional disaster.

Appendix A — Tools Reference

Reverse Image Search

- Google Lens — lens.google.com — broadest coverage
- Yandex Images — yandex.com/images — best for faces
- TinEye — tineye.com — best for finding the OLDEST occurrence (sort by date)
- Bing Visual Search — bing.com/visualsearch — strong on Western content

Image Forensics

- Forensically — 29a.ch/photo-forensics — ELA, clone detection, noise analysis
- InVID WeVerify — weverify.eu — keyframe extraction and Deepfake tab
- FotoForensics — fotoforensics.com — quick ELA

Video Analysis

- InVID WeVerify Deepfake Detection tab — ensemble of 5 neural networks
- Deepware Scanner — deepware.ai — face manipulation specialist
- YouTube frame-by-frame — press the period (.) key during playback

Audio Analysis

- Sonic Visualiser — sonicvisualiser.org — free spectrogram viewer
- Audacity — audacityteam.org — free audio editor with spectrum view
- ElevenLabs AI Speech Classifier — elevenlabs.io — only reliable on unedited ElevenLabs output

Geolocation

- Google Earth Pro — free desktop application with street view
- SunCalc — suncalc.org — calculate sun position by date and location
- Wikimapia — wikimapia.org — annotated satellite maps

Metadata Inspection

- ExifTool — command-line, comprehensive metadata extraction
- Jeffrey's Image Metadata Viewer — exif.regex.info/exif.cgi — browser-based

Appendix B — Signal Quick-Reference Cards

Print these. Keep them at your desk. Use them under deadline pressure when memory fails.

TEXT — 8 Signal Categories

53. Vocabulary: delve, foster, showcase, harness, profound, intricate, tapestry, realm, represents, moving beyond
54. Triplets: groups of three (THE MASTER KEY)
55. Reversal structures: 'not just X, it's Y'
56. Modular sentences: shuffle test passes, few connectors
57. Missing hedging: no 'I think', 'maybe', 'perhaps'
58. Sophistry: 50+ word paragraphs that summarise in 6 words
59. Structural defaults: 800-1000 words, 4-6 paragraphs, em dashes
60. Tonal markers: bland confidence, no personal voice

IMAGE — 8 Inspection Categories

61. Hands and anatomy: count fingers, check teeth, eyes, ears
62. Lighting: single source, consistent shadow direction
63. Edges: hair against background, glasses, jawlines
64. Text: read every visible word, look for distortion
65. Skin texture: too smooth, too symmetric
66. Background: repeating patterns, ghostly figures
67. Context: physics, scale, time period, cultural plausibility
68. Provenance: ✦ watermark, filename, EXIF

VIDEO — 10 Inspection Categories

69. Temporal consistency (THE MASTER KEY): static objects flicker?
70. Faces and lip sync: mouth shapes match sounds?
71. Hands and anatomy in motion
72. Physics: gravity, inertia, cloth, fluids
73. Lighting across time: shadows track motion correctly?
74. Background stability: people, signs, geometry
75. Audio-visual sync: footsteps, ambience
76. Resolution and quality patterns
77. Multi-shot coherence (if cuts present)
78. Provenance markers: watermarks, platform signatures

VOICE — 8 Inspection Categories

79. Breathing and biology (THE MASTER KEY): can you hear breaths?

80. Pitch and prosody: too smooth?
81. Pauses and rhythm: any fillers, ums, restarts?
82. Articulation: robotic precision or natural slurring?
83. Emotional authenticity: matches content?
84. Drift over time: changes across the clip?
85. Background environment: silent or naturally ambient?
86. Technical markers: spectral signatures, ElevenLabs polish

A Final Word

You will not catch every fake. No one will. The technology improves faster than detection. By the time you finish reading this book, some of the specific signals it teaches will be obsolete — fixed by the next generation of generative models. The vocabulary watchlist will need updating in two years. The hand-counting trick is already less reliable than it was in 2023. New signatures will replace the old ones.

What will not change is the methodology. Layered verification. Multiple independent checks. Calibrated scepticism. Honest acknowledgement of uncertainty. The discipline to choose AMBER when the evidence does not support GREEN. The humility to revise your judgment when new information arrives.

Tools change. Methodology endures. The five-step protocol in Chapter 6 will outlive every specific signal in this book.

If you remember nothing else, remember this: the goal of verification is not to catch fakes. The goal is to make better decisions. Sometimes that means catching a deepfake. Sometimes it means defending a real recording from the Liar's Dividend. Sometimes it means publicly admitting you do not know.

Train your eyes. Trust your methodology. Stay sceptical. Stay humble. Stay committed to the truth.

— Umut İlhan

Valmera AI Consulting & Training

April 2026