DETECTING AI IMAGES IN CRIMINAL DEFENSE CASES

Practical Forensic Tools for Evaluating Image Evidence

Kenneth G. Hartman

GSE, CISSP, GCFA, GCFE, GASF

Lucid Truth Technologies | Digital Forensics Consultant

Welcome everyone. Today we're going to talk about something that is increasingly affecting criminal defense work: AI-generated images and how to evaluate whether image evidence in your cases is authentic, synthetic, or simply unverifiable. I'm going to give you practical tools — things you can use in discovery motions, cross-examination, and expert retention — not just theory.

AGENDA

01 **The Democratization of Deception**
How AI images threaten evidence integrity

02 **Hygiene & Metadata Analysis**
First-line checks any attorney can request

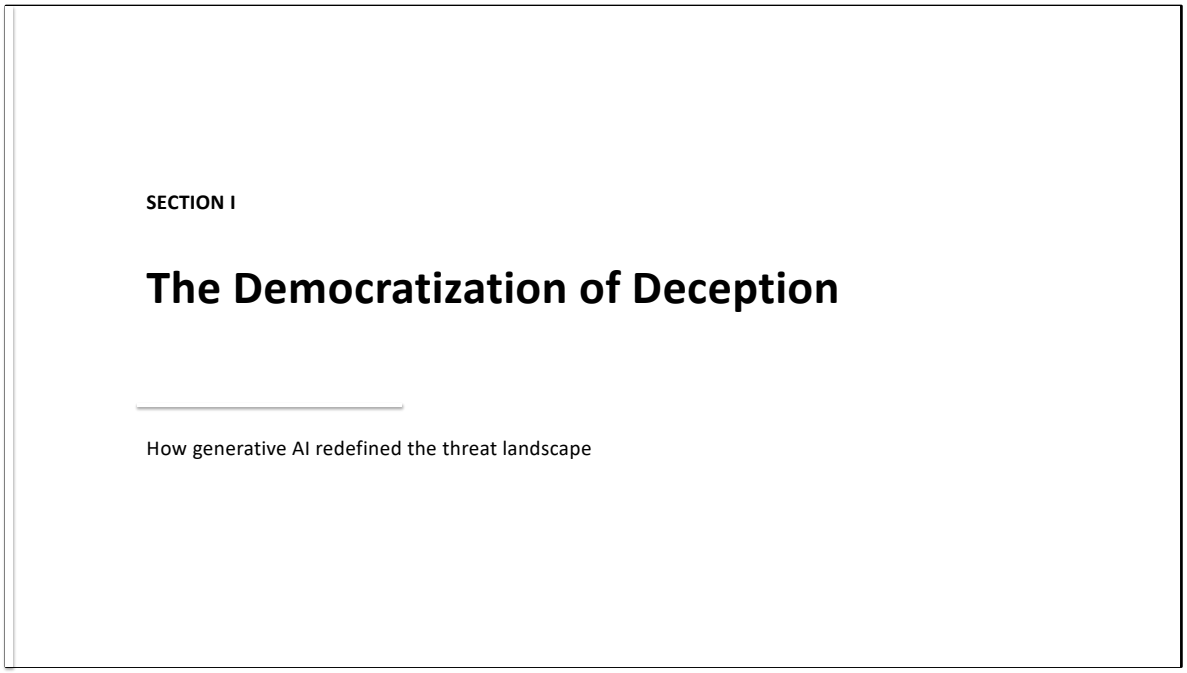
03 **Camera Ballistics — The "Gold Standard"**
Court-admissible source identification

04 **Detecting the Ghost in the Machine**
AI detection tools & their legal limits

05 **The Future of Authentication (C2PA)**
Cryptographic provenance & what it means for evidence

06 **Discovery, Motions & Practical Takeaways**
Actionable strategies for your next case

Here's our roadmap. We'll start with why AI-generated images matter to your practice right now. Then we'll work through four layers of forensic analysis — from checks you can do yourself all the way to techniques that produce court-admissible evidence. We'll finish with a practical discovery checklist you can use in your next case where image authenticity is at issue.



[Transition slide — pause briefly]

This first section sets the stage. Before we talk about solutions, we need to understand just how dramatically the problem has shifted in the last few years.

The Current Landscape

- **Early Deepfakes:** Face-swaps and crude manipulations gave way to sophisticated neural synthesis
- **Generative Adversarial Networks (GANs):** Generator vs. discriminator produces photorealistic output
- **Diffusion Models:** DALL·E, Midjourney, Stable Diffusion — text-to-image at unprecedented quality

Each generation of models makes previous detection methods less reliable

Five years ago, deepfakes were mostly face-swaps in videos — impressive but detectable. Today, we have full text-to-image generation that produces photorealistic output from a text prompt. The key shift is from GANs — where a generator and discriminator compete to produce realistic images — to diffusion models like DALL-E, Midjourney, and Stable Diffusion. Diffusion models work by learning to remove noise from images, and they've surpassed GANs in quality and diversity. The critical point: every generation of these models makes the previous generation's detection methods less reliable. This is an arms race.

The Threat Vector

- **Accessibility:** Open-source tools lower the barrier — anyone with a GPU can generate convincing fakes
- **Fraud & Evidence Tampering:** Synthetic images used in insurance fraud, legal disputes, identity theft
- **Disinformation:** Fabricated images of events, people, and documents spread at scale
- **The Liar's Dividend:** Real evidence dismissed as "probably AI-generated" — doubt weaponized

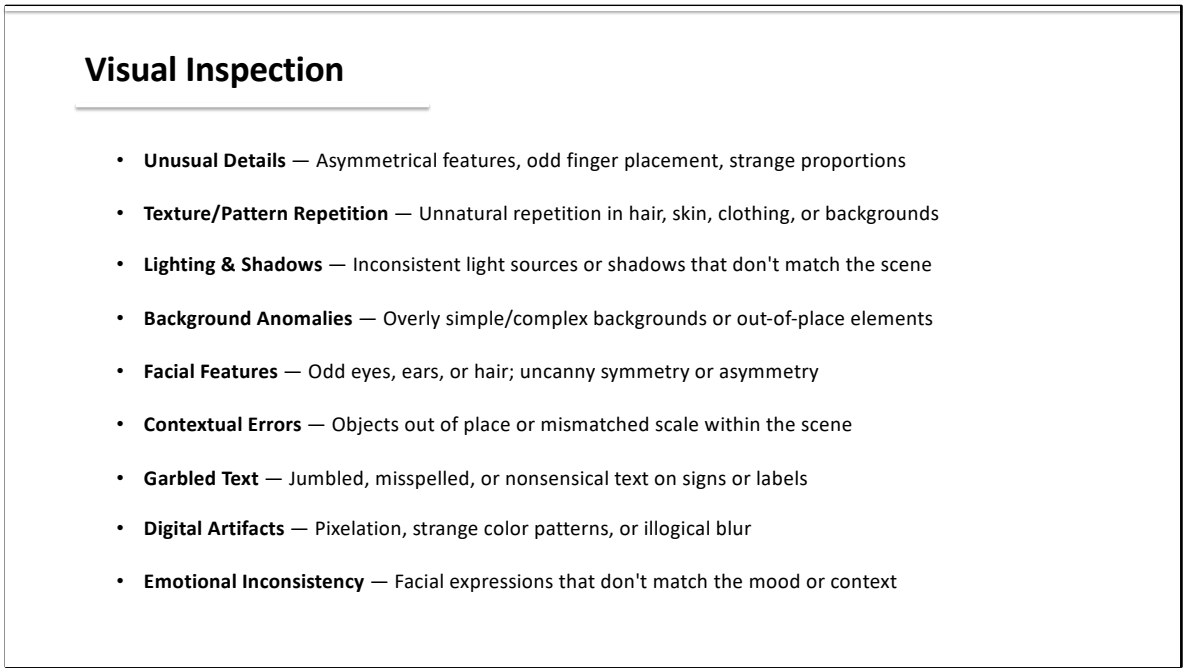
https://www.etsy.com/listing/1667241073/realistic-silicone-sixth-finger-prop

The democratization angle is what makes this urgent. You don't need a PhD in machine learning anymore. Open-source models run on consumer GPUs. That means the barrier to creating convincing fake images has collapsed. We're seeing this in insurance fraud — fabricated damage photos. In legal disputes — manufactured evidence. In identity theft — synthetic ID documents. But perhaps the most insidious threat is what researchers call the "liar's dividend": once people know fakes are possible, **real** evidence gets dismissed as "probably AI-generated." The mere existence of the technology undermines trust in all digital media.

The Liar's Dividend

- **The Concept —** Coined by law professors Chesney & Citron (2019): deepfakes don't just enable fabrication — they give bad actors a plausible excuse to deny authentic evidence
- **The Inversion —** The danger is no longer just fake evidence being believed; it's real evidence being dismissed
- **In the Courtroom —** A defendant can claim any genuine photo, video, or audio was AI-generated — and create reasonable doubt without proving it
- **The Erosion Effect —** As deepfake awareness grows among juries, so does generalized skepticism toward all digital evidence
- **Already Happening —** Defense attorneys and defendants have invoked "that could be AI" in response to legitimate evidence in real cases
- **The Paradox —** The better deepfake detection becomes, the more confidently bad actors deny authentic evidence knowing detection is imperfect
- **The Stakes —** Victims of real crimes may see their genuine evidence dismissed; guilty parties may escape accountability behind manufactured doubt

Five years ago, deepfakes were mostly face-swaps in videos — impressive but detectable. Today, we have full text-to-image generation that produces photorealistic output from a text prompt. The key shift is from GANs — where a generator and discriminator compete to produce realistic images — to diffusion models like DALL-E, Midjourney, and Stable Diffusion. Diffusion models work by learning to remove noise from images, and they've surpassed GANs in quality and diversity. The critical point: every generation of these models makes the previous generation's detection methods less reliable. This is an arms race.



Here is a practical checklist they can apply immediately — no software required. Walk through each indicator briefly, but don't rush. These are the signals that a trained eye learns to spot.

Unusual Details — Start here. AI generators struggle with fine motor anatomy. Hands are the classic tell — count the fingers. Look for ears that don't match, or clothing that blends into the background in odd ways.
Texture and Pattern Repetition — AI "tiles" complex textures. Fabric, hair, and brick walls are frequent failure points. Look for that copy-paste quality in areas that should have natural variation.
Lighting and Shadows — This one requires a little training but pays off. Ask: where is the light source? Now check whether every object in the scene is lit from that same direction. Shadows that fall the wrong way are a strong indicator.
Background Anomalies — Attorneys often focus on the subject and miss the background entirely. Train yourself to look at the edges and corners. AI often gets lazy there.
Facial Features — Eyes are the most common failure point — look at the reflections in the pupils, the shape of the iris, and whether both eyes are actually the same. Hair is another giveaway, especially at the hairline.

Contextual Errors — Step back and ask: does this scene make sense? A coffee cup that's twice the size of someone's head, or a window that opens onto an impossible view — these are context failures the AI didn't catch.
Garbled Text — This is one of the easiest checks. If there's a sign, label, or text anywhere in the image, zoom in. AI consistently fails at readable, coherent text.
Digital Artifacts — Look for blur that doesn't belong — a sharp face on a blurry neck, or pixelation at boundaries between objects. These are seams where the generation failed.
Emotional Inconsistency — This is subtle but powerful in front of a jury. A face that is technically correct but feels *wrong* — the smile doesn't reach the eyes, the expression doesn't match the situation. Trust that instinct and investigate further.
Key message: None of these alone is conclusive — but two or three together are a strong basis for requesting expert examination. This checklist is your triage tool.

Test Your Skill

Image Grid from:
How to Distinguish AI-Generated Images from Authentic Photographs
Negar Kamali, Karyn Nakamura, Angelos Chatzimparmpas, Jessica Hullman, Matthew Groh

https://arxiv.org/abs/2406.08651

Drop it into Gemini:

Look at each of the images in this screen shot and let me know which ones are fake and which ones are real based on your best judgment, and explain why you made the determination.

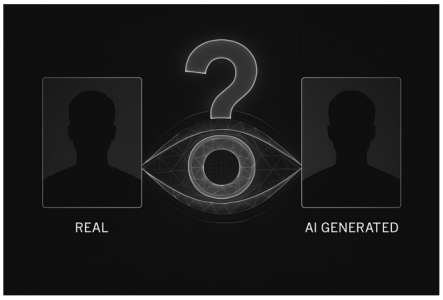
Top Left (Man walking): Look closely at the hand holding the briefcase; the fingers are fused and malformed. Additionally, the architectural ribs of the tunnel merge confusingly in the background, lacking logical structural geometry.
Top Middle (Doctor): The primary failure point here is the stethoscope. The acoustic tubing on his right side vanishes directly into his shirt rather than connecting logically to the earpieces.
Top Right (Soldier): The chin strap of the headgear merges seamlessly into the subject's skin and beard. The vehicle in the background also consists of amorphous, undefined shapes rather than mechanical parts.
Middle Left (Man with dog): There are severe anatomical anomalies in the man's hand resting on the dog. The fingers are elongated, broken, and fuse together unnaturally. The dog's ear also blends directly into the man's coat.

Middle Center (Woman in park): While one of the cleaner generations, it exhibits the hyper-flawless skin texture and perfect lighting characteristic of modern diffusion models. Structurally, the collar of her grey shirt on the right side ends abruptly and asymmetrically.
Middle Right (Three women): The glasses on the center woman have a thick frame on one side that simply disappears on the other. The hand of the woman on the far right is a blurred, amorphous shape lacking distinct fingers.
Bottom Left (Women in green): The hand and arm of the woman in the foreground merge unnaturally into the folds of the fabric, forming a claw-like shape. The face of the person partially visible on the far left edge is also heavily distorted.
Bottom Middle (Elderly couple): On both subjects, the temple pieces of their glasses vanish into thin air before reaching their ears. The background subject's eye structure also shows significant generation bleed.
Bottom Right (Men hugging): This panel shows classic limb generation failures. The hand resting on the back has malformed, indistinguishable fingers, and the hand on the waist melts directly into the suit fabric. The ties also disappear into the shirts without logical layering.

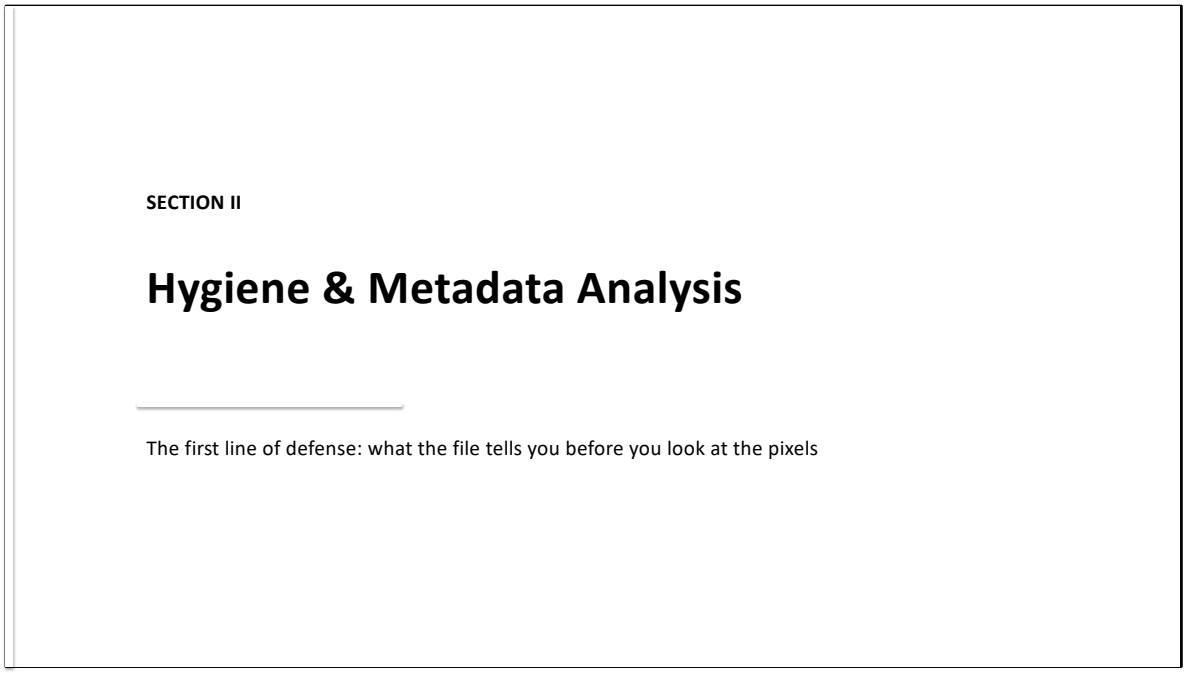
The Challenge

- State-of-the-art generators pass casual human review >90% of the time
- A layered forensic approach is required — no single technique is enough
- We need tools that examine what humans cannot see

Human visual inspection is no longer sufficient.



This is the key takeaway from the introduction. I want this to sink in: human visual inspection is no longer sufficient. Studies show that state-of-the-art generators fool casual human reviewers over 90% of the time. Even trained analysts struggle with the latest models. We need instruments — forensic tools that examine things the human eye literally cannot see. That's what the rest of this presentation is about: building a layered toolkit.



[Transition slide]

Layer one is what I call "hygiene." It's the fastest, cheapest check — and also the most easily defeated. But it's still worth doing because it catches the low-hanging fruit, and it's where every investigation should start.

Exif & File Signatures

Layer 1: What the metadata reveals

- **Software Tags:** Generative tools often leave telltale "Software" metadata (e.g., "DALL·E", "Stable Diffusion")
- **File Header Anomalies:** Synthetic images may have unusual compression profiles or missing camera-specific fields
- **Tool Spotlight — ExifTool:** Industry-standard CLI for extracting and analyzing all embedded metadata
  - **Quick wins:** Check Make, Model, Software, GPS, and date/time fields for inconsistencies
  - **Red flags:** Missing EXIF entirely, or camera fields present but no lens data

When an image comes in for examination, the first thing I do is run ExifTool. It's the Swiss Army knife of metadata extraction. You're looking for a few things: Does the "Software" tag reference a generative tool? DALL-E, Stable Diffusion, and some others leave fingerprints here. Does the file have camera-specific fields — Make, Model, lens data, GPS? A real photograph from a real camera has a rich metadata profile. A generated image might have none, or might have camera fields but no lens data — which is suspicious. The red flag isn't just missing metadata; it's inconsistent metadata. A file that claims to be from a Canon EOS R5 but has no lens information and a compression profile that doesn't match Canon's JPEG engine — that tells a story.

ExifTool

```
Aperture                        : 1.9
Image Size                      : 4080x3072
Megapixels                      : 12.5
Scale Factor To 35 mm Equivalent: 5.4
Shutter Speed                   : 1/492
Create Date                     : 2026:01:04 13:09:21.964-05:00
Date/Time Original              : 2026:01:04 13:09:21.964-05:00
Modify Date                     : 2026:01:04 13:09:21.964-05:00
Thumbnail Image                 : (Binary data 26410 bytes, use -b option to extract)
GPS Altitude                    : 194.4 m Above Sea Level
GPS Date/Time                   : 2026:01:04 18:08:56Z
GPS Latitude                    : 46 deg 34' 30.52" N
GPS Longitude                   : 85 deg 15' 20.71" W
MP Image 2                      : (Binary data 9722 bytes, use -b option to extract)
Circle Of Confusion             : 0.006 mm
Depth Of Field                  : inf (0.46 m - inf)
Field Of View                   : 112.6 deg
Focal Length                    : 2.2 mm (35 mm equivalent: 12.0 mm)
GPS Position                    : 46 deg 34' 30.52" N, 85 deg 15' 20.71" W
Hyperfocal Distance             : 0.46 m
Light Value                     : 11.8
Lens ID                         : Pixel 8 Pro back camera 2.23mm f/1.95
```

Limitations of Metadata Analysis

Metadata can be stripped, forged, or simply absent.

- Social media platforms strip EXIF data on upload — most shared images have no metadata
- Sophisticated actors deliberately craft fake metadata to mislead investigators
- Metadata presence proves nothing by itself — it must be corroborated
- We need techniques that analyze the image content itself, not just the wrapper

https://lucidtruthtechnologies.com/photo-metadata/

Here's the honest caveat. Metadata analysis is necessary but not sufficient. Social media platforms strip EXIF data on upload, so most images you encounter in the wild have no metadata at all. Sophisticated actors can craft fake metadata — it's just XML and binary fields, trivially editable. And the presence of legitimate-looking metadata doesn't prove authenticity — it only proves someone put it there. This is why we need deeper layers. We need to analyze the image content itself, not just its wrapper.

Visual Forensics

Layer 1: Pixel-level inspection techniques

- **Lighting Consistency:** Check if light sources, highlights, and shadows are physically coherent across the scene
- **Error Level Analysis (ELA):** Re-compress at known quality and compare — manipulated regions show different error levels
- **Compression Artifacts:** JPEG quantization patterns differ between original captures and composites
- **Perspective & Geometry:** Vanishing points, reflections, and proportions that violate physical reality

Beyond metadata, we can look at the pixels themselves. Lighting consistency is a big one — in a real photograph, there's a coherent set of light sources that produce consistent shadows and highlights. Composites and some AI-generated images get this subtly wrong. Error Level Analysis — ELA — is a technique where you re-save the image at a known JPEG quality and then look at the difference. Regions that have been modified or inserted from a different source show different error levels than the rest of the image. It's not foolproof, but it's a useful triage tool. You can also check perspective geometry — do vanishing points converge correctly? Are reflections physically consistent?

Error Level Analysis

Detecting Compression Anomalies in Digital Images

What Is ELA?

- Re-saves the image at a known JPEG quality level
- Computes the pixel-level *difference* between re-saved and original
- Amplifies that difference into a visible map
- Uniform regions = consistent compression history (authentic)
- Bright/hot regions = higher error = different origin or recent modification

What to Look For:

🔴 **Hot edges around objects** → composite insertion

🔴 **Inconsistent error levels** between foreground and background → different source materials

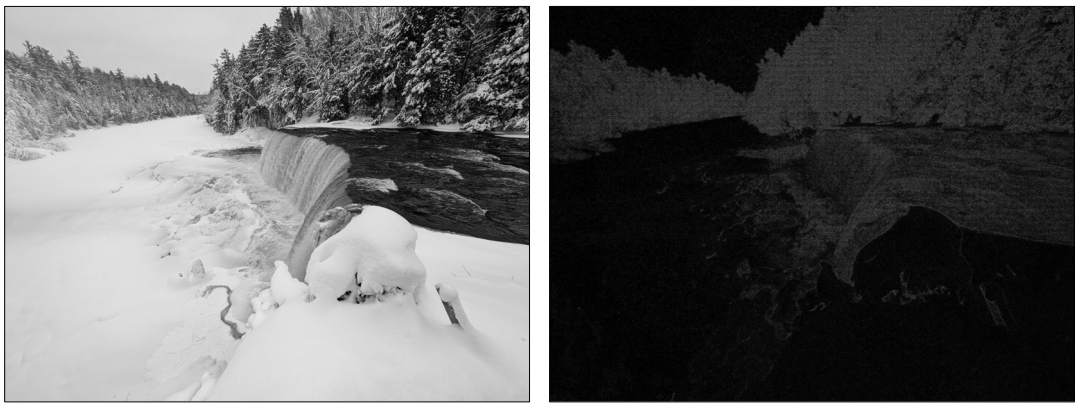
🔴 **Uniform darkness across entire image** → heavily recompressed (provenance concern)

🟢 **Consistent, low-level texture throughout** → authentic single-capture image

https://29a.ch/photo-forensics/#error-level-analysis

Here's the underlying logic. Every time a JPEG image is saved, it loses a small amount of information to compression. If an image was captured in one shot and saved once, the entire image has a consistent compression history — it all "ages" together. When you apply ELA, the error map should be relatively uniform across the image.

ELA – Known Good Image (Pixel)



This is your baseline — what ELA looks like when it's working exactly as expected on an authentic, single-capture photograph.

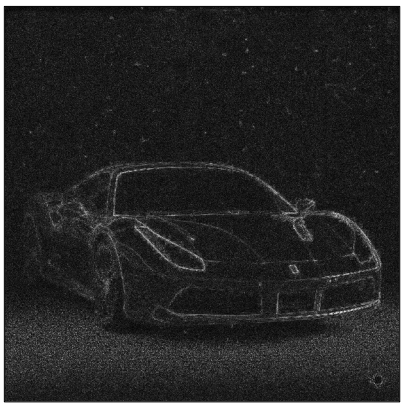
The image on the left is a real photograph taken with a Pixel phone — a winter waterfall scene. Notice the ELA output on the right. It's not perfectly black, and that's normal and expected. What you're seeing is the natural texture of a genuine photograph. Areas with high contrast and fine detail — the churning water, the snow-covered branches, the edge of the falls — show some error activity because those regions have genuine high-frequency content that compresses differently than smooth areas. The dark sky and flat snow fields are nearly black because they're uniform and compress very cleanly.

The key thing to train your eye on is consistency. The error levels track the image content in a logical way — detail shows more error, smooth areas show less. There are no hard, unnatural edges. There are no regions that seem to belong to a different image. The whole scene "breathes" together as a single captured moment.

This is what you want to see when you run ELA on evidence you're trying to

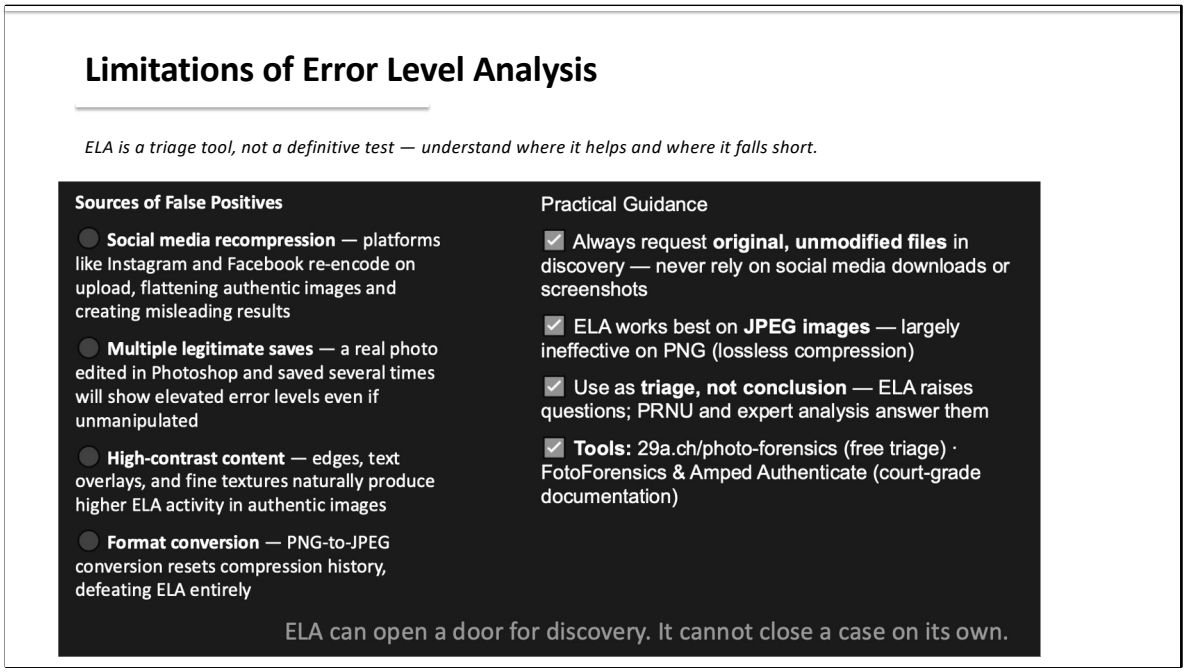
authenticate. Keep this reference image in mind as we look at the next slide, because the contrast is going to be stark.

ELA – Nano Banana Image



Now here's where it gets forensically interesting. If someone takes a photo of a car and composites it into a different background scene — or if an AI generator assembles an image from components with different statistical origins — those regions carry *different* compression histories. They haven't aged together. When ELA amplifies the difference, the inserted element lights up like a spotlight while the surrounding scene stays dark.

Look at the example on this slide. The car's body panels, roofline, and front fascia are all "hot" — orange and magenta in the ELA output. The road, trees, and background houses are nearly black. And critically, that hard luminous edge exactly tracing the car's boundary is the tell. That's not a natural artifact of a single-capture photograph. That's the seam of a composite.



A few practical caveats your audience should understand. ELA is a triage tool, not a definitive test. Social media recompression can create false positives — platforms like Instagram and Facebook re-encode images on upload, which can flatten authentic content and create misleading ELA results. That's why you always want the original file, not a screenshot or a social media download. And ELA works best on JPEG images — it's less effective on PNGs, which use lossless compression.

The tool I'd recommend for quick triage is 29a.ch/photo-forensics — it's free, browser-based, and requires no installation. For more rigorous forensic work, FotoForensics and Amped Authenticate provide more controls and documentation suitable for expert witness reports.

The takeaway for defense attorneys: if prosecution produces an image and you can demonstrate through ELA that the subject and background have different compression histories, you have grounds to challenge authenticity under Rule 901 and to request the original unmodified file in discovery.

AI Watermarking: Invisible Signatures from the Source

When the Generator Marks Its Own Output

- **Google SynthID —** Imperceptible pixel-level watermark; survives cropping, resizing, screenshots, and print/re-photograph cycles
- **Google Visible Watermark —** Four-pointed star icon in the bottom-right corner of Gemini-generated images; trivially removed with any photo editor
- **Meta Invisible Watermarking —** Frequency-domain embedding on Imagine-generated images; detectable after heavy editing
- **Meta Stable Signature —** Watermark baked into the model decoder itself; every output pre-watermarked at the architecture level
- **OpenAI / DALL·E 3 —** No pixel-level watermark; relies on metadata disclosure only
- **The Critical Gap —** Open-source models (Stable Diffusion, Flux, ComfyUI) watermark nothing — and that's exactly where evidentiary risk lives

Before we get into camera ballistics, let me briefly address what some of you are probably thinking: "Can't AI companies just mark their own images?" Yes — some are trying. But it's more complicated than it sounds.

Google SynthID — Invisible Watermark -- SynthID embeds an imperceptible pattern directly into pixel values at generation time — below human perception but detectable by Google's algorithm. Impressively resilient: survives cropping, resizing, color adjustments, even print and re-photograph cycles. The catch: only works on Google's own models, and detection requires their proprietary detector. You cannot run it yourself.

Google Visible Watermark -- Gemini-generated images also get a small four-pointed star icon in the bottom-right corner — a human-readable disclosure for casual viewers. The limitation: trivially removed with any photo editor. This is disclosure, not detection. Know the difference.

Meta Invisible Watermarking -- Meta embeds watermarks in the frequency domain of images from Imagine and Meta AI — similar conceptually to what we'll examine shortly in frequency analysis. Claims to survive heavy editing and

compression.

Meta Stable Signature -- The most technically interesting approach: watermark baked directly into the model decoder itself. Every output is pre-watermarked at the architecture level — no post-processing step to bypass. You'd have to retrain the entire model to remove it.

OpenAI / DALL·E 3 -- No pixel-level watermarking. Relies entirely on metadata disclosure — covered in the C2PA section later.

The Critical Gap -- Here's the takeaway: all of this is voluntary, proprietary, and platform-specific. The images appearing in your cases aren't coming from Google or Meta's tools — they're coming from Stable Diffusion, Flux, and ComfyUI running on someone's laptop. Those tools watermark nothing, log nothing, and leave no trace by design. There's also an active research community building watermark removal tools. This is an arms race. Watermarking is promising long-term — but not a reliable forensic tool today.

** So we can't count on the generator to identify itself. Let's talk about what we CAN rely on — the physics of the camera that captured a real image — and a technique called camera ballistics.

SECTION III

Camera Ballistics — The "Gold Standard"

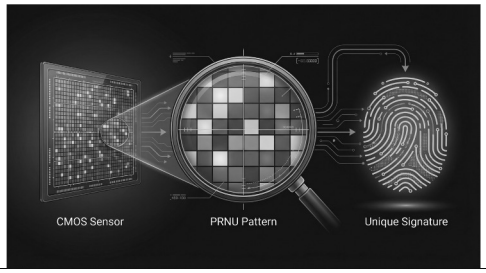
Using sensor fingerprints to authenticate the origin of an image

This is where forensic science gets really interesting, and where we move into court-admissible territory. Camera ballistics is the gold standard for image source identification, and it works on principles borrowed from firearms forensics.

Sensor Fingerprinting: PRNU

Layer 2: The "digital fingerprint" of a camera sensor

- **Photo Response Non-Uniformity:** Unique "digital bullet scratches" from manufacturing imperfections in the sensor
- **Every Camera Is Unique:** Even identical make/model cameras produce different PRNU patterns
- **Persistent & Involuntary:** The pattern is embedded in every photo the sensor captures — it cannot be turned off
- **Analogy:** Like matching a bullet to a specific gun barrel — hence "camera ballistics"

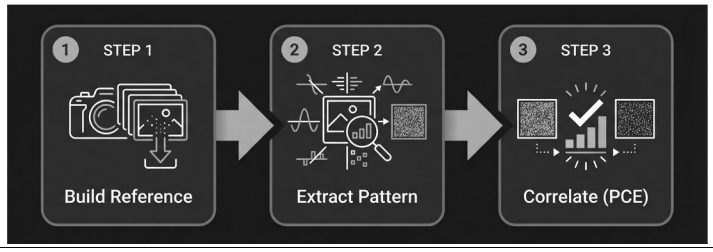


Every digital camera sensor has manufacturing imperfections at the pixel level. These imperfections cause each pixel to respond slightly differently to the same amount of light. This variation is called Photo Response Non-Uniformity — PRNU. Think of it exactly like the scratches inside a gun barrel: every barrel is unique, and every bullet fired through it picks up that unique pattern. Every camera sensor is unique, and every photo taken with it picks up that unique noise pattern. It's persistent — it doesn't change over time. It's involuntary — you can't turn it off. And it's unique — even two cameras of the same make and model have different PRNU signatures. This is the "digital fingerprint" of a specific physical device.

Forensic Validity of PRNU

Layer 2: Why courts accept this evidence

- **Peer-Reviewed Science:** Decades of published research supporting PRNU-based identification
- **Daubert Standard:** Meets the legal threshold for admissibility of scientific evidence in U.S. courts
- **Court-Tested:** Successfully used in criminal cases for source identification
- **Key Application:** Proving an image did NOT come from a specific camera flags it as potentially synthetic
- **Limitation:** Requires access to the suspected source camera (or images from it) to build the reference fingerprint



This is critical for anyone working in legal contexts. PRNU-based identification has decades of peer-reviewed research behind it. It meets the Daubert standard — the legal threshold for admissibility of scientific evidence in U.S. federal courts. It has been successfully used in criminal prosecutions. For our purposes — detecting deepfakes — the key application is proving that an image did NOT come from a claimed camera. If someone says "I took this photo with my iPhone," and the PRNU analysis shows no correlation with that iPhone's sensor, you have strong forensic evidence of fabrication. The main limitation: you need either the actual suspect camera or a set of images known to have been taken with it. You can't do PRNU analysis in a vacuum.

SECTION IV

Detecting the Ghost in the Machine

AI-specific forensic techniques that identify synthetic origins

Now we move from camera-focused forensics to AI-focused forensics. These techniques don't look for the fingerprint of a camera — they look for the fingerprint of a generator.

Frequency Domain Analysis

Layer 3: What the frequency spectrum reveals

- **Fourier Transform:** Convert spatial image data to frequency domain to reveal hidden patterns
- **GAN Artifacts:** Generators often produce unnatural grid patterns or periodic artifacts invisible to the naked eye
- **High-Frequency Signatures:** Real cameras produce characteristic noise profiles — synthetic images do not match
- **Spectral Anomalies:** Checkerboard patterns in the frequency domain are a strong indicator of GAN-generated content

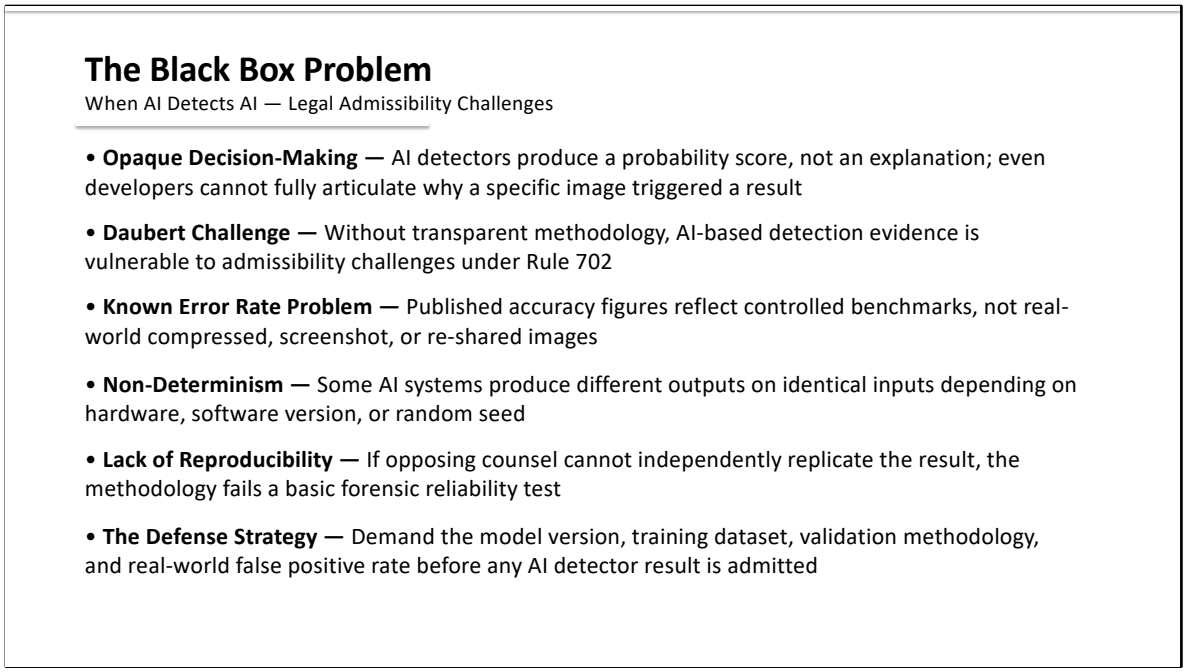
This is one of my favorite techniques because it's so elegant. When you apply a Fourier transform to a real photograph, the frequency spectrum has a natural, organic distribution — it reflects the physics of light, optics, and sensor capture. When you do the same to a GAN-generated image, you often see something completely different: unnatural grid patterns, periodic artifacts, or checkerboard patterns that are invisible in the spatial domain but light up like a Christmas tree in the frequency domain. These artifacts come from the upsampling layers in the generator's architecture. Diffusion models produce their own distinct frequency signatures related to the denoising process. The key insight: the frequency domain reveals what the pixel domain hides.

Automated Detectors & Attribution

Layer 3: Scaling detection with machine learning

- **CLIP-Based Models:** Vision-language approaches that generalize across unseen generators
- **Model Attribution:** Classifying images to specific tools (Midjourney v5 vs. DALL·E 3 vs. Stable Diffusion)
- **Transfer Learning:** Detectors trained on known generators can partially generalize to new ones
- **Ensemble Approaches:** Combining multiple detection signals improves accuracy and reduces false positives
- **Challenge:** New model architectures require continuous retraining of detectors

The detection field is moving toward scalable, automated solutions. CLIP-based models — which combine vision and language understanding — have shown promising ability to generalize across generators they weren't explicitly trained on. Ensemble approaches that combine multiple detection signals reduce false positives. But I want to be honest about the limitation: every time a new model architecture appears, existing detectors need retraining. This is an inherent challenge — we're always one generation behind the generators. That's precisely why we need the layered approach, not reliance on any single detection method.



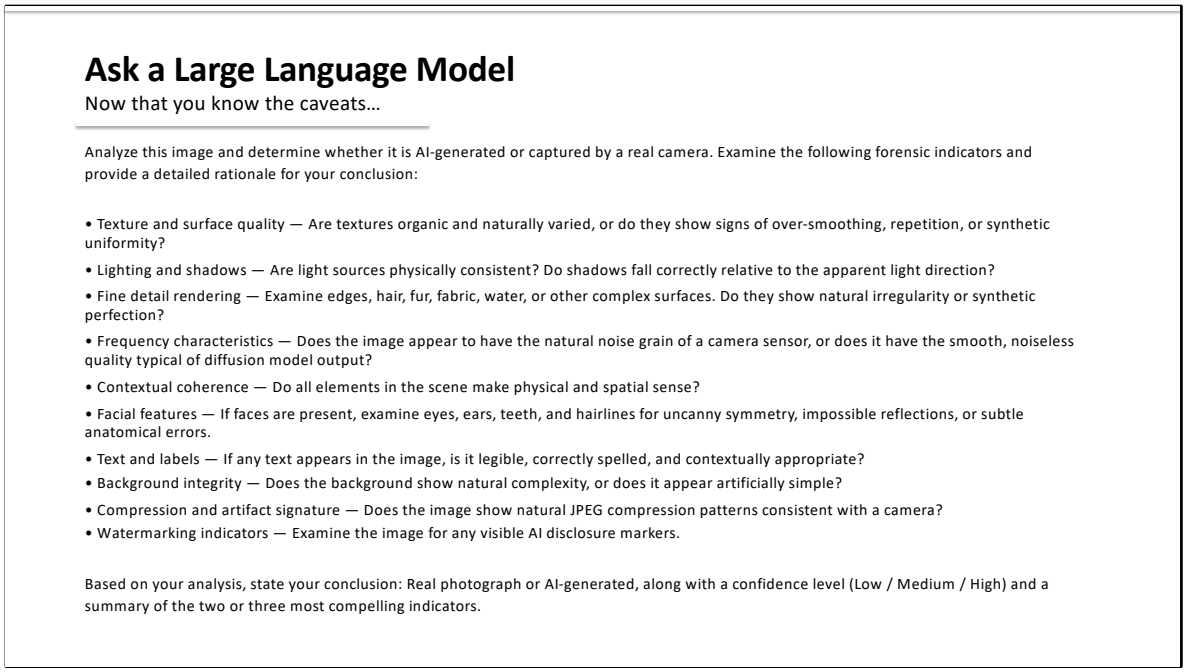
The Black Box Problem — Legal Admissibility of AI Detectors
This is where I want your attorney ears to perk up, because this is directly relevant to how you challenge or defend AI-based detection evidence in court. When an AI detector tells you an image is synthetic, it is not giving you a reason — it's giving you an output. The internal decision-making process is opaque. The model examined millions of pixel relationships through dozens of neural network layers and produced a probability score. Nobody — including the tool's developers — can fully explain why that specific image triggered a positive result. That's the black box problem.

This creates serious challenges under Daubert and Rule 702. For expert testimony relying on AI detection tools to be admissible, the methodology must be testable, peer-reviewed, have a known error rate, and be generally accepted in the relevant scientific community. AI detectors, particularly newer ones, struggle on several of those prongs:
Known error rate — Most published accuracy figures come from controlled benchmark datasets. Real-world false positive and false negative rates on the kinds of compressed, screenshot, and re-shared images that appear in criminal cases are substantially worse and often unpublished.

Testability — If you can't explain how the model reached its conclusion, opposing counsel can argue the methodology is not independently verifiable.
Non-determinism — Some AI systems produce slightly different outputs on the same input depending on hardware, software version, or random seed. That inconsistency is difficult to reconcile with the reproducibility requirement courts expect of forensic methods.
Transparency — Unlike PRNU, which has decades of published mathematical foundations you can point to, a neural network classifier is not easily reduced to a formula a jury can evaluate.
The practical implication for defense attorneys: if the prosecution is relying on an AI detector as evidence that an image is real, demand the methodology. What model? What version? What was the training dataset? What is the documented false positive rate on images of this type? Has it been independently validated? Those are the questions that expose the weakness.

And the flip side: if you are retaining an expert who uses AI detection tools to challenge evidence, make sure they are using those tools as one layer of a documented, transparent, multi-method analysis — not as a standalone conclusion. An expert who says "the AI said it's fake" will not survive cross-examination. An expert who says "the AI flagged it, which I then corroborated through frequency analysis and PRNU mismatch" is on much stronger ground.

Transition: "This is exactly why no single tool is sufficient — and why the layered approach isn't just a best practice, it's a legal necessity."



The Black Box Problem — Legal Admissibility of AI Detectors
This is where I want your attorney ears to perk up, because this is directly relevant to how you challenge or defend AI-based detection evidence in court. When an AI detector tells you an image is synthetic, it is not giving you a reason — it's giving you an output. The internal decision-making process is opaque. The model examined millions of pixel relationships through dozens of neural network layers and produced a probability score. Nobody — including the tool's developers — can fully explain why that specific image triggered a positive result. That's the black box problem.

This creates serious challenges under Daubert and Rule 702. For expert testimony relying on AI detection tools to be admissible, the methodology must be testable, peer-reviewed, have a known error rate, and be generally accepted in the relevant scientific community. AI detectors, particularly newer ones, struggle on several of those prongs:
Known error rate — Most published accuracy figures come from controlled benchmark datasets. Real-world false positive and false negative rates on the kinds of compressed, screenshot, and re-shared images that appear in criminal cases are substantially worse and often unpublished.

Testability — If you can't explain how the model reached its conclusion, opposing counsel can argue the methodology is not independently verifiable.
Non-determinism — Some AI systems produce slightly different outputs on the same input depending on hardware, software version, or random seed. That inconsistency is difficult to reconcile with the reproducibility requirement courts expect of forensic methods.
Transparency — Unlike PRNU, which has decades of published mathematical foundations you can point to, a neural network classifier is not easily reduced to a formula a jury can evaluate.
The practical implication for defense attorneys: if the prosecution is relying on an AI detector as evidence that an image is real, demand the methodology. What model? What version? What was the training dataset? What is the documented false positive rate on images of this type? Has it been independently validated? Those are the questions that expose the weakness.

And the flip side: if you are retaining an expert who uses AI detection tools to challenge evidence, make sure they are using those tools as one layer of a documented, transparent, multi-method analysis — not as a standalone conclusion. An expert who says "the AI said it's fake" will not survive cross-examination. An expert who says "the AI flagged it, which I then corroborated through frequency analysis and PRNU mismatch" is on much stronger ground.

Transition: "This is exactly why no single tool is sufficient — and why the layered approach isn't just a best practice, it's a legal necessity."

SECTION V

The Future of Authentication (C2PA)

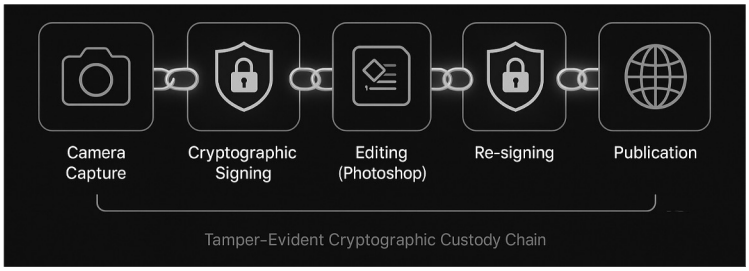
From detecting fakes to proving reality

Everything we've discussed so far is reactive — analyzing an image after the fact to determine if it's real. C2PA flips the model entirely. Instead of detecting fakes, we prove reality.

Content Credentials (C2PA)

Shift the paradigm: prove an image is REAL rather than trying to prove it is FAKE.

- **Coalition for Content Provenance and Authenticity:** Industry standard for signed metadata
- **Cryptographic Signatures:** Tamper-evident chain of custody from capture to publication
- **Manifest System:** Records every edit, export, and transformation with cryptographic proof



This is a paradigm shift. The Coalition for Content Provenance and Authenticity — C2PA — is an industry standard backed by Adobe, Microsoft, Intel, Nikon, and others. The idea: embed cryptographically signed metadata into an image at the moment of creation, and maintain that chain through every edit, export, and transformation. If the chain is intact and the signatures verify, you know the provenance of the image. If the chain is broken or absent, you can't make provenance claims — but you still have our other layers to fall back on. Think of it like chain of custody for physical evidence, but implemented with cryptography.

Hardware vs. Software Trust

Camera-Embedded

- Credentials signed at moment of capture
- Hardware-rooted trust (TPM/secure enclave)
- Unbroken chain of custody from sensor to file
- **Manufacturers:** Leica, Sony, Nikon
- Highest security assurance

Software-Added

- "Digital labels" added post-creation
- **Applications:** Photoshop, Firefly, DALL·E
- Offers transparency about edits and AI use
- Lower security assurance — no hardware root of trust
- Still valuable for disclosure and chain of edits

This distinction is critical and often overlooked. Camera-embedded credentials — from Leica, Sony, Nikon — are signed at the moment of capture using hardware-rooted trust, typically a TPM or secure enclave on the camera itself. The chain of custody starts at the sensor. This is the highest assurance level. Software-added credentials — from Photoshop, Firefly, or AI generation tools — are "digital labels" added after creation. They're valuable for transparency: they tell you "this image was edited in Photoshop" or "this was generated by DALL-E." But they don't have a hardware root of trust, so the security assurance is lower. Both are useful. But in a courtroom, hardware-signed credentials from the capture device carry significantly more weight.

Discovery & Motions Checklist

- **Request Original Files** — Motion to compel production of original, unmodified image files — not screenshots, printouts, or social media downloads
- **Demand Device Access** — If image is claimed to be a photograph, request access to (or images from) the alleged source device for PRNU analysis
- **Challenge Metadata** — File EXIF data showing "Software: DALL·E" or missing camera fields entirely — basis for 901(b)(9) challenge
- **Request Expert Examination** — If State relies on image evidence, demand independent forensic examination under Rule 706 or retain your own expert
- **Challenge AI Detector Evidence** — If prosecution uses an AI detection tool, demand model version, training data, error rates, and reproducibility under Daubert/Rule 702
- **Invoke the Liar's Dividend Defense** — If your client's image evidence is dismissed as "probably AI," demonstrate that this skepticism must apply equally to the State's exhibits
- **Check C2PA Credentials** — If image claims provenance via Content Credentials, verify whether credentials are hardware-signed (high assurance) or software-added (lower assurance)

This is your practical takeaway slide. Each of these items maps to a specific motion or discovery strategy. The key principle: the earlier you challenge image evidence, the stronger your position. Don't wait for trial to raise authenticity concerns — raise them in discovery, in pretrial motions, and in Daubert hearings. The prosecution bears the burden of authentication under Rule 901, and these tools give you specific, articulable bases to challenge their exhibits.

SECTION VI

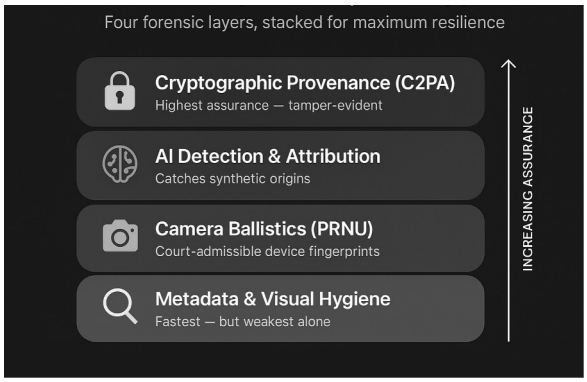
Conclusion & Practical Takeaways

Assembling your forensic toolkit

Let's bring it all together.

Defense in Depth: A Layered Approach

- **Layer 1 — Metadata & Visual Hygiene:** Quick triage with ExifTool and ELA (fast but bypassable)
- **Layer 2 — Camera Ballistics (PRNU):** Court-admissible source identification (requires reference camera)
- **Layer 3 — AI Detection & Attribution:** Frequency analysis, model fingerprinting, CLIP classifiers
- **Layer 4 — Cryptographic Provenance (C2PA):** Hardware-signed credentials prove authenticity at capture
- **No Single Layer Is Sufficient:** Combine multiple techniques for robust verification



This is the framework I want you to take away. Layer 1 — metadata and visual hygiene — is your quick triage. Fast but bypassable. Layer 2 — camera ballistics with PRNU — is your court-admissible heavy artillery, but it requires a reference camera. Layer 3 — AI-specific detection through frequency analysis and model fingerprinting — catches what PRNU can't and gives you attribution. Layer 4 — C2PA cryptographic provenance — is the long-term solution, proving authenticity from the moment of capture. No single layer catches everything. Use them together. Start with metadata, escalate to PRNU if you have the reference device, run AI detection in parallel, and check for C2PA credentials if they exist.

Key Takeaways

- **Visual Inspection Is Your Starting Point:** Train yourself to spot AI artifacts — hands, text, lighting inconsistencies — before requesting expert analysis
- **Discovery Is Your Most Powerful Tool:** Original files, source devices, and metadata are more valuable than any detection algorithm
- **Know When to Retain an Expert:** PRNU analysis meets Daubert and can definitively link or exclude a camera — use it when device attribution matters
- **AI Detectors Are Vulnerable to Challenge:** Black-box tools with unknown error rates face serious admissibility hurdles under Rule 702
- **C2PA Changes the Game:** Hardware-signed credentials create verifiable chains of custody — their absence doesn't prove fakery, but their presence proves provenance

Five things to remember. One: this is a cat-and-mouse game, and it always will be. Generators will adapt to evade detection. Two: layered defense is non-negotiable — no single tool is enough. Three: PRNU meets the Daubert standard and C2PA creates auditable chains — both are court-relevant today. Four: invest in C2PA infrastructure — hardware-signed credentials are the strongest long-term play. Five: stay current — both the detection tools and the generative models are evolving rapidly. Continuous learning isn't optional in this field.

QUESTIONS & DISCUSSION

Discovery Strategy • Evidence Admissibility • Expert Retention

Five things to take back to your practice. One: train your eye — visual inspection catches the obvious cases and tells you when to dig deeper. Two: discovery is everything — always demand original files and source devices. Three: PRNU is your heavy artillery — it's the only technique with decades of Daubert-tested peer review. Four: if the prosecution is using an AI detector, challenge it hard — these tools have serious admissibility vulnerabilities. Five: C2PA is coming — hardware-signed credentials will change how we authenticate images, and you need to understand the difference between hardware-signed and software-added.

ABOUT THE SPEAKER

Kenneth G. Hartman

GSE, CISSP, GCFA, GCFE, GASF

- **Founder, Lucid Truth Technologies:** Digital forensics consultancy specializing in contested digital evidence for criminal defense attorneys throughout Michigan
- **Licensed Private Investigator:** State of Michigan License No. 3701207402; legally authorized to conduct forensic investigations and generate court-admissible work product
- **GIAC Security Expert (GSE #198):** Fewer than 250 holders worldwide; the most rigorous certification in information security
- **SANS Certified Instructor:** Currently teaching SEC502: Cloud Security Tactical Defense; prior security leadership at Google, Illumina, and SAP Ariba
- **Published Researcher:** Three-part blog series on AI image authentication; GIAC Gold papers on Amazon EC2 forensics and BitTorrent investigations; SANS Reading Room contributor since 2014
- **Criminal Defense Specialist:** CDITC certified; Associate member of CDAM, Michigan Council of Professional Investigators, and National Association of Legal Investigators

Contact me at ken@lucid-truth.com for a copy of my most recent CV.