GUIDE · AUGUST 30, 2026

Artificial Intelligence

A Technical Guide for Lawyers and Judges

By Law & Forensics LLC

Download the PDF (65 pages)

KEY TAKEAWAYS

- “AI” names at least four different kinds of system — rule-based automation, machine-learning classifiers, generative models, and agentic systems — which fail differently and leave different records.
- A generative model predicts likely continuations rather than retrieving facts, which is why it can produce fluent, well-formed, entirely false content without anything having malfunctioned.
- When AI is used in a matter, the discoverable material is almost always an operational record — prompts, tool calls, retrieved document identifiers, outputs, and timestamps — not the model itself, and most of it sits with a third-party provider on a short retention clock.
- Rule 901(a) requires only evidence sufficient to support a finding that an item is what its proponent claims; the 2026 “it's a deepfake” objection is structurally identical to the 2011 “my account was hacked” objection, which the existing rules already absorbed.

- There is no generally accepted error rate for synthetic-media detection, which makes the Daubert error-rate factor the natural pressure point on any detection expert.
- Requiring a specific and articulable basis before an authenticity inquiry opens — stated in the first case-management order, before any dispute arises — is the most effective single tool a court has in this area.
- Rule 37(e) reserves its severest sanctions to a finding of intent to deprive; and the absence of a prompt log is usually evidence that logging was never enabled, not that anything was destroyed.
- Sanctions in the AI cases have followed the failure to verify, not the use of the tool: AI does not alter a lawyer's preexisting professional duties.

What This Guide Is, and What It Is Not

This guide is written for lawyers, judges, magistrate judges, arbitrators, and law professors who need to reason accurately about artificial intelligence in a case — not for computer scientists. It assumes a reader who is comfortable with the Federal Rules of Evidence and the Federal Rules of Civil Procedure and uncomfortable being told that a system is a black box and left there.

It is organized the way a technical subject has to be organized for a legal audience: bottom-up. Part I explains what the technology is and how it behaves. Part II explains what it produces and leaves behind — the artifacts that can actually be preserved, produced, and examined. Part III applies the existing rules of evidence and procedure to those artifacts. Part IV turns the whole thing into questions counsel can ask at a Rule 26(f) conference and questions a court can decide at a first case-management conference. Each part depends on the one before it. A reader who skips Part I and starts at the deepfake chapter will find the vocabulary unfamiliar and the reasoning thin, because the reasoning genuinely does rest on the mechanics.

Two expectations are worth setting aside now.

This guide does not tell you whether AI is good or bad. That question is not useful in a courtroom. The useful questions are narrower and answerable: what did this system do, what record did it leave, what can that record support, and what error characteristics does it have. A judge does not need a position on artificial intelligence. A judge needs to know whether a particular exhibit is what its proponent says it is, and what evidence bears on that.

This guide does not contain a detection recipe. There is no reliable checklist that separates a synthetic video from a genuine one by inspection, and any source that offers one is selling something. What exists instead is a set of investigative steps — provenance at source, native files and metadata, chain of custody, examination

by a qualified examiner using stated tools — that shifts the question from intuition to evidence. That is less satisfying and considerably more durable.

A word about currency. The rulemaking process is deliberate; the interval between a proposal and an adopted rule is conventionally measured in years. Generative systems change on a scale of months. That mismatch is not an accident to be corrected but a structural feature of the situation, and it is one of the stated reasons the Advisory Committee on Evidence Rules has repeatedly preferred to test whether existing rules can absorb a new problem before writing a rule that may be obsolete on adoption. Practically, it means two things for the reader: the existing rules are doing more work in this area than one might expect, and any statement in this guide about the current state of a proposal, a standard, or a regulation should be checked against the primary source before it is relied on. Where this guide is aware that something is in motion, it says so rather than freezing a snapshot into an assertion.

Finally, nothing here is legal advice, and nothing here substitutes for a qualified forensic examiner in a contested matter. The purpose of the guide is to make counsel and the court harder to mislead — by an overconfident expert, by an opponent's rhetoric, or by their own intuitions about machines.

KEY TAKEAWAYS

- This guide is descriptive, not predictive: it explains mechanisms and the questions those mechanisms raise, not how any particular dispute should come out.
- The technology described here changes faster than the rules that govern it, which is one reason the Advisory Committee on Evidence Rules has so far preferred to apply existing rules rather than write new ones.
- Nothing here is a detection recipe, and nothing here substitutes for a qualified examiner in a contested matter.

Vocabulary: What “AI” Names, and What It Does Not

The single most common source of confusion in AI-related litigation is that the phrase artificial intelligence names at least four different kinds of system that behave differently, fail differently, and leave different records. Precision here is not pedantry; it determines what can be requested, what can be preserved, and what an expert can honestly say.

Rule-based automation. A program that applies conditions a human wrote: if the invoice exceeds this amount, route it for approval. Nothing is learned. Its behavior is fully specified by its source and configuration, and it is reproducible — the same inputs produce the same output every time. Much of what is marketed as AI is this. It is also the easiest category to examine, because the rules can be read.

Machine learning, in the classical supervised sense. A system that is shown examples labeled by humans and derives, from those examples, a function that assigns labels to new items. Spam filters, fraud scoring, and technology-assisted review in discovery are all of this type. The important properties for a lawyer are that its accuracy is a measurable quantity, that its behavior depends on the examples it was trained on as much as on its design, and that a bias in the training examples propagates into the outputs in a way that is invisible from the outputs alone.

Generative models, including the large language models behind current chat assistants and the image, audio, and video generators behind synthetic media. These are trained on very large collections of text, images, or audio and learn statistical relationships within that material. Given a prompt, the system produces new content that is statistically consistent with the patterns it learned. It is not retrieving a stored answer and it is not reasoning to a conclusion in the way a person does; it is generating an output that resembles what such an output usually

looks like. That distinction is the origin of nearly every evidentiary problem in this guide.

Agentic systems. A generative model wired to act: it pursues an assigned goal, calls external tools, queries databases, and executes multi-step workflows with limited human supervision. The legally significant fact about an agent is that it takes actions in systems of record, which means it leaves an operational trail — and that the trail, not the model, is usually the discoverable thing.

Three further distinctions are worth carrying.

Training and inference are separate phases. Training is the expensive, one-time (or periodically repeated) process that produces a model from data. Inference is what happens when a user submits a prompt and gets an output. Records from the two phases sit in different places, are held by different parties, and are subject to different retention. A request framed as "all data used by the AI" collapses the two and will usually produce either nothing or an objection. A request framed around inference-time records — prompts, retrieved documents, tool calls, outputs, timestamps — is answerable.

Artificial general intelligence is not what is deployed. The systems that appear in litigation are narrow in the sense that matters: they perform tasks within a domain and have no understanding of the world, no goals of their own, and no capacity to verify their own claims. Arguments that proceed from what a sufficiently intelligent machine might do are not arguments about the exhibit in front of the court.

The words are not synonyms. An *algorithm* is a procedure. A *model* is the specific trained artifact — in current systems, a large set of numeric parameters (often called weights) produced by training. A *tool* is the product a vendor sells. A *system* is the tool plus the surrounding configuration, data sources, and controls. A *deployment* is one organization's specific installation of that system, with its own settings, access rules, logging, and retention. Almost every discovery dispute that gets framed as a demand for "the algorithm" is really a demand about a

deployment, and reframing it that way is usually what makes it proportionate and answerable.

A handful of technical terms recur and are worth defining once, in plain terms. A **token** is the unit a language model reads and writes — roughly a word fragment. An **embedding** is a numeric representation of a piece of text or an image that places semantically similar items near one another, and is what makes retrieval systems work. A **context window** is the amount of material a model can consider at once; anything outside it is simply not before the system. **Fine-tuning** is additional training applied to a general model to specialize it. **Retrieval-augmented generation** is the common enterprise pattern in which the system searches a document collection and places the retrieved passages into the model's context before generating an answer — which is why retrieval records matter evidentially: they show what the model was actually shown. **Multimodal** describes a system that accepts or produces more than one kind of content, such as text and images together.

KEY TAKEAWAYS

- "AI" names at least four different families of system — rule-based automation, machine-learning classifiers, generative models, and agentic systems — which behave differently and leave different records.
- Training and inference are distinct phases; a discovery request that does not distinguish them will not reach what the requesting party actually needs.
- "Algorithm," "model," "tool," "system," and "deployment" are not interchangeable words in a discovery request, and the difference between them is usually the difference between a producible record and an unanswerable demand.

How a Generative Model Produces an Output

Courts already have a workable mental model for a learning system, and it comes from discovery rather than from computer science. In technology-assisted review, an attorney codes a modest set of documents as responsive or not. The system identifies properties of those documents and uses them to code the rest, and the process repeats until the system's predictions and the reviewer's coding sufficiently coincide. Judges have been accepting that explanation for more than a decade. It is a good on-ramp because it makes three things concrete: the system learns from examples rather than from instructions, its output is a prediction rather than a determination, and its quality is something you measure rather than something you assume.

A generative model extends that idea in scale and in kind. During training it is exposed to an enormous body of text (or images, or audio) and adjusts its internal parameters so that it becomes very good at one narrow task: predicting what comes next. For text, that means predicting the next token given everything before it. The same principle applies to pixels in an image or to samples in a piece of audio. What the model acquires is not a store of facts but a very detailed statistical picture of how such material is usually put together.

At inference time, the model is given a prompt and produces a continuation, one token at a time, each token chosen from a probability distribution over the vocabulary. Two consequences follow directly, and both matter in court.

The system can produce fluent, well-formed, entirely false content without anything having gone wrong. If the pattern in the training data is that a legal brief cites cases in a particular format, the model will produce something in that format when the context calls for it, whether or not a matching case exists. Nothing in the mechanism checks the world. This is commonly called hallucination, which is an unfortunate term because it suggests a malfunction. It is better understood as the predictable behavior of a system optimized to produce plausible continuations

rather than true ones. It is not a bug that will be patched away, and a party arguing that a fabricated citation was an aberration is arguing against the design of the tool.

Output is not deterministic. The selection of each token typically involves sampling rather than always taking the single most probable option, and the degree of randomness is usually a configurable setting. The practical result is that the same prompt, submitted twice to the same system, can produce different answers. Systems are also updated: a provider may change the underlying model, its configuration, or its safety filters without notice to the user, so a result obtained in March may not be obtainable in September even with identical inputs.

That second consequence is the point at which the technology collides directly with forensic practice. Digital forensics has long treated reproducibility as an admissibility floor: a report should permit an independent third party, given the same inputs, to replicate the conclusion. A step in an examination workflow that produces a different answer on a second run does not meet that expectation on its own terms. This does not make AI-assisted analysis inadmissible, but it does mean that where such a step is used, the examiner must record enough to make the run reconstructable — the system and version, the configuration, the exact prompt, the material supplied, the output received, and the date — and must be prepared to explain what a second run would and would not be expected to reproduce.

Three related points close out the mechanics.

First, a model has no privileged access to its own reasoning. When a system produces an explanation of why it answered as it did, that explanation is itself generated text, produced by the same predictive process. It may be accurate; it is not a log, and it should never be treated as one.

Second, memorization and generalization are different. Models generalize from patterns, but they can also reproduce distinctive passages from their training data,

which is the technical root of several ongoing copyright disputes. The relevant point for evidence is that the presence of a passage in an output does not establish where it came from.

Third, and most practically: **“the model said so” is not a foundation.** An assertion produced by a generative system is hearsay-adjacent in the colloquial sense and evidentially inert in the technical one — it is not a person's statement, not a business record, and not the output of a process shown to produce a reliable result. What can be admissible is evidence about the system: that it was run, on what inputs, producing what output, recorded how. That is an entirely different exhibit, and it is the one counsel should be building.

KEY TAKEAWAYS

- A generative model predicts likely continuations; it does not look up facts, which is why it can produce fluent, well-formed, entirely false content without any malfunction occurring.
- Because generation is stochastic, the same prompt can produce different outputs on different runs — a property in direct tension with the forensic expectation that a result can be reproduced.
- Fluency is a property of the output, not evidence of its truth, and “the model said so” is not a foundation for anything.

Variability, Error, and the Limits of the Technology

Every claim about what an AI system can do is a claim about a rate, and rates are measured, not asserted. The vocabulary for measuring them already exists in discovery practice and transfers cleanly.

Recall is the share of all genuinely responsive items that the process actually found. **Precision** is the share of the items the process returned that are genuinely responsive. They trade off: a broader net catches more of what you want and more of what you do not. A tool tuned to miss nothing will flag a great deal that is innocuous, and a tool tuned to flag only what it is confident about will quietly miss things. A vendor claim of high accuracy that does not say which of these it is measuring, on what population, is not information. (These two definitions are transposed with some regularity in practitioner literature, in both directions. Check any formula you are handed against the plain-language definitions above rather than assuming the printed version is right.)

Richness — the proportion of the underlying population that is genuinely responsive — is a property of the collection, not of the tool, and it drives everything else. A very low richness population makes high precision arithmetically difficult and makes recall expensive to measure, which is why sampling plans in low-richness collections need to be designed rather than improvised.

The **null set** is where a flawed process hides its failures: the material the process did not return. Nobody looks there by default, so nothing corrects the error. The standard remedy is an elusion test — draw a random sample from the items the process rejected, review them, and estimate from that sample how much responsive material the process missed. This is the only routine way recall is actually measured rather than assumed, and it is inexpensive relative to what it protects against. The governing principle is short enough to remember: a search

methodology with no validation is an assertion of completeness with no evidence behind it.

Two error characteristics are specific to the AI setting and have no clean analogue in older technology.

Detection performance is structurally unstable. Many generative systems are trained in a loop against a discriminator whose job is to tell generated content from real content; improving the generator's ability to defeat a detector is, in that architecture, the training objective. Even where a modern generator is not built that way, the commercial and adversarial dynamics run in the same direction: detectors are published, generators are updated, and the detector's measured accuracy on last year's outputs tells you progressively less about this year's. A detection tool's error rate is therefore not a fixed property but a snapshot with a short half-life, and an expert offering a detection opinion should be asked when the tool was last validated and against what.

Detection cues that were once reliable have degraded. Early forensic discussion of synthetic video emphasized physiological tells — irregular blinking, unnatural pulse signals in facial video, subtle expression mismatches — on the theory that generated content fails to mimic complex human behavior. Those cues were already weakening several years ago. Presenting them today as current detection practice is a mistake, and an expert who relies on them without addressing how current generators handle them has a problem on cross-examination.

The honest summary of the state of the art, and it is important that a guide of this kind say it plainly: **there is no generally accepted error rate for synthetic-media detection.** Consensus on what constitutes a reliable test, and on what an acceptable error rate would be, is still developing. Detection models require continuous updating to remain useful at all. This is not a criticism of the researchers working on the problem; it is the state of a young field. But it has a direct doctrinal consequence, developed further below: the *Daubert* factor asking for the known or potential rate of error is not a formality in this area. It is the place

where a detection opinion is most vulnerable, and a court that presses on it is asking the right question.

Set against this is a requirement that predates AI entirely and is not going to relax to accommodate it. Forensic conclusions are expected to be reproducible: a report should let an independent examiner, given the same evidence, reach the same result, and a report offered without the underlying forensic images is a recognized warning sign. Where conclusions cannot be reproduced, they deserve correspondingly little weight. Holding a non-deterministic process to a reproducibility standard is the most productive tension in this entire subject, and it is not resolved by asserting that the technology is new. It is resolved by documentation: recording enough about each run that an independent examiner can evaluate what was done, even where they cannot obtain a byte-identical repetition.

KEY TAKEAWAYS

- Accuracy is a measured property of a specific tool on a specific population, never a vendor adjective — and recall and precision trade off against each other.
- A search or classification methodology with no validation is an assertion of completeness with no evidence behind it.
- There is no generally accepted error rate for synthetic-media detection, which makes the Daubert error-rate factor the natural pressure point on any detection expert.

The Artifacts an AI System Leaves Behind

The question a litigator actually has is not how a transformer works. It is: a party used an AI system in this matter — what exists now that I can preserve, request, and examine? The answer is more concrete than the subject's reputation suggests, and it is almost never the model.

Inference-time records. A well-instrumented deployment logs, for each interaction, the prompt submitted, any tool the system called and with what arguments, the identifier of each document the system retrieved and placed in its context, the output returned, and the timestamp of each. That set is what makes after-the-fact reconstruction possible. It answers the questions that actually matter — what was the system asked, what was it shown, what did it do, what did it say, and when — without requiring anyone to open the model. Whether a particular deployment captures all of it is exactly the question to ask early, because the answer is a configuration choice and configuration choices are usually documented.

Retrieval and access records. In the common enterprise pattern, the system searches a document collection and feeds what it finds into the model. Controls at that retrieval layer — role-based and attribute-based access rules, partitioning of the vector store, and allowlists governing which tools the system may call — are both the practical security mechanism and a rich evidentiary source, because they define and record what the system was permitted to reach. One consequence deserves emphasis for anyone assessing what an organization knew: a capability change, such as adding a new tool or a new data source, can expand an agent's effective reach without altering its nominal permissions. The permission table alone will not show it. The tool-call log will.

The deployment record. Independent of any single interaction, a governed AI deployment generates documents: a written AI risk assessment; a system or model card describing the system's intended purpose and known limitations;

evidence of evaluation, red-teaming, and adversarial testing; minutes of whatever committee approved the deployment; a named accountable owner; and an incident-response plan. Where the legal question is what the organization knew or had reason to know about how its system could behave, this record is the evidence — often more directly than any log. Counsel on both sides should note the privilege dimension early: risk assessments prepared at the direction of counsel may be privileged, the same document prepared as ordinary compliance work usually is not, and the distinction is made at the time the document is created, not at the time it is requested.

Outputs and their lineage. The output of an AI system rarely stays where it was produced. It is pasted into a draft, summarized into an email, or committed to a system of record, and by the time it becomes an exhibit, the fact that a machine produced it may be invisible on the face of the document. Establishing lineage — this paragraph came from that session, on that date — is usually a matter of correlating timestamps and content across the interaction log and the document's own history, which is why both need to be preserved together.

Two practical cautions.

First, **most of these records sit with a third-party provider, not with the party.** Enterprise AI is generally consumed as a hosted service. Retention of prompts and logs is a contractual and configuration matter, the default windows are often short, and some enterprise configurations are explicitly zero-retention, meaning the provider keeps nothing after the session. Counsel who assume that a log exists because logs usually exist will find out otherwise, and will find out late. The retention terms and the deployment's configuration should be established at the very beginning of the matter, in writing, and the preservation demand should be specific enough that the provider's own retention clock is not the thing that decides the case.

Second, **incident response should preserve more than the incident.** Where an AI system is implicated in a loss or a dispute, the material worth capturing

immediately includes the prompt logs, the tool-use logs, the documents that were retrieved, and, where it is feasible at all, the model artifacts themselves — meaning a record of exactly which model version and configuration were in service. Model versions change. A system that gives one answer in March and a different one in September is not evidence of tampering; it is evidence that nobody recorded which version was running.

KEY TAKEAWAYS

- When AI is used, the discoverable material is almost always an operational record — prompts, tool calls, retrieved document identifiers, outputs, and timestamps — rather than the model itself.
- A well-governed deployment generates a documentary record (risk assessment, system or model card, evaluation and red-team results, governance minutes, named owners) that is frequently more probative than anything inside the model.
- Most of these records sit with a third-party provider on contractual retention terms, which is why the preservation clock on AI evidence runs faster than counsel expect.

The Substrate: Storage, Modification, Retention, and Deletion

Before a lawyer can reason about AI artifacts, they need an accurate picture of how digital information is stored, changed, and lost. This is the least glamorous part of the subject and the part that most often decides motions.

A storage device is divided into sectors, which the file system groups into clusters — the smallest unit it will allocate to a file. A file larger than one cluster occupies several, and those clusters need not be adjacent; a file scattered across non-contiguous clusters is described as fragmented. The file system maintains a table recording which clusters belong to which file, and whether each file is considered active.

Deletion, in ordinary use, does not erase anything. Emptying the recycle bin typically flips a flag: the file is marked as no longer in use, and its clusters are marked as available for reallocation. The bytes are still on the disk. Until the operating system actually writes something else into those clusters, the file remains recoverable in its original form. Once some clusters have been reused, the file is at best partially recoverable, and often not recoverable at all. **Recoverability is therefore a function of time and use, not of intent** — which is why the interval between the deletion and the preservation of the device is often the single most consequential fact in a spoliation motion.

The corollary is one of the most frequently misused facts in ESI practice, and it cuts against the party alleging spoliation more often than counsel expect. Overwriting is continuous and mostly not user-driven. Operating-system updates, indexing, antivirus scans, defragmentation, backup software, and ordinary application activity consume free space constantly, at a rate measured in thousands of clusters per second on an active system. **The mere absence of a document is insufficient evidence that anyone wiped it.** An examiner who says a file is gone has said something about the current state of the disk, not about

anyone's conduct, and the inferential distance between the two has to be argued rather than assumed.

The distinction between allocated and unallocated space is worth an analogy, because it explains why proportionality arguments about deep forensic recovery so often succeed. Allocated space is like documents filed in labeled cabinets: organized, indexed, and searchable. Unallocated space is like unbundled piles of shredded paper. Both may contain relevant material; the cost of extracting it differs by orders of magnitude, and courts weigh that difference under Rule 26(b)(1).

Metadata comes in at least two kinds that behave very differently. System metadata is maintained by the operating system and file system — created, modified, and accessed timestamps, file size, location. Application metadata is maintained by the program that created the file — author, revision history, tracked changes, comments, embedded properties. The two can disagree, and where they do, the disagreement is itself evidence. Critically for practice, **metadata is generally not reproduced when a document is printed to paper or converted to a static image**, which is why form of production is not a housekeeping detail: producing a native file and producing a PDF of the same document produce materially different exhibits.

Merely opening evidence can spoil it. Access timestamps update on open. Documents containing automatic date fields update those fields when the file is opened in the originating application, so the date on the face of the document silently becomes today's date. This is why examination proceeds from a forensic image rather than from the original: a bit-for-bit image, taken through a write blocker so the acquisition process cannot itself modify the source, captures the whole medium including deleted and unallocated space, and is verified with a cryptographic hash so that any later copy can be proven identical. A hash is a one-way function producing a fixed-length digest; a change of a single bit changes the digest completely, which is what makes it useful as an integrity check.

Practitioners should note that MD5, which appears throughout older forensic literature as the default, is no longer collision-resistant and SHA-256 is the current expectation; where a report relies on MD5 alone, that is worth a question, though in most non-adversarial contexts it remains adequate for detecting accidental corruption.

Two modern realities complicate the classical account above, and any source that does not mention them is dated. **Solid-state storage** behaves differently: the TRIM command and the drive's internal garbage collection can erase the contents of deleted blocks proactively, without any user action, which substantially undercuts the assumption that deleted data persists until overwritten by new writes. And **encryption is now the default** on mainstream phones and laptops, which changes every acquisition decision — an image of an encrypted volume without the key is an image of noise.

Now the AI overlay, and it is stark. **There is no bit-for-bit forensic image of a hosted AI service.** No one is going to acquire a provider's model, and even if they could, the model's parameters would not answer any question a court is asking. The substrate for AI evidence is a log in someone else's system, governed by that provider's retention policy and by the customer's configuration. Everything the classical account teaches about time, use, and loss applies with more urgency, not less: provider retention windows are commonly measured in days or weeks rather than years, session context is discarded when the session ends, and the underlying model can be replaced beneath the evidence without anyone being notified. The same under-collection problem that already affects ephemeral and quasi-ephemeral messaging channels — which are routinely missed not through concealment but because standard collection workflows simply do not reach them — extends directly to AI chat interfaces, and for exactly the same reason.

KEY TAKEAWAYS

- Deleting a file ordinarily marks its space as available rather than erasing its contents, so recoverability is a function of time and subsequent use — not of intent.
- The absence of a document is not, by itself, evidence that someone destroyed it; ordinary system activity reclaims space continuously and without user involvement.
- There is no bit-for-bit forensic image of a hosted AI service, which is why AI evidence is almost entirely log-derived and why provider retention terms are load-bearing.

A Taxonomy of AI-Touched Evidence

The phrase AI evidence is used to cover at least five distinct things, which behave differently under the rules and call for different discovery. Separating them is the most useful single move available to counsel at the outset of a matter.

(i) Content generated by an AI system. Text, a document, an email, a report, a summary. The authentication question is ordinary — who created this, when, from what — but the answer often requires establishing that a machine, not the ostensible author, produced the language. The discovery targets are the interaction records described above. The evidentiary risk is fabricated content presented as researched content, which is the failure mode behind the sanctions cases.

(ii) Authentic content altered by AI. A real contract with modified terms, a real log file with edited entries, a real photograph with an element removed or added. **This is the hardest category and receives the least attention.** A wholly fabricated document invites scrutiny of the whole; a genuine document with one altered paragraph carries its own corroboration, because everything around the alteration checks out. The examination is a comparison problem — against a counterpart copy, against the surrounding system records, against the document's own internal consistency — rather than a detection problem.

(iii) Wholly synthetic audio, video, or images. Deepfakes in the ordinary sense. These differ in kind from historical manipulation because they do not merely alter a depicted event; they can fabricate the event entirely, and generated elements can be combined seamlessly with genuine footage so that the exhibit is partly real. This is the category with the most public attention and, as discussed below, the one where the doctrinal question is less about detection than about who bears what burden and when.

(iv) AI-generated or AI-suggested metadata. This category is real and is commonly described inaccurately, so it deserves care. A language model does not

populate a file's hash value — a hash is computed by the examining tool over the file's bytes and cannot be forged by a system that writes text. What actually happens is more mundane and more dangerous: a person accepts a model-suggested author name, title, or date and enters it into document properties; or a template carries stale properties forward into a new document; or an automated pipeline stamps fields with values derived from generated content. **Forensic tools report what a file contains, not whether the contents are truthful.** A properties pane showing a creation date of 2019 on a brief written in 2025, or a privilege log whose document properties name a reviewer who does not exist, will be reported faithfully by the tool and can send an inquiry in an entirely wrong direction. The remedy is cross-checking against independent records — mail system logs, document management system history, custodian testimony — rather than trusting any single field.

(v) The AI system's own operational records. The prompts, tool calls, retrieval records, outputs, model and system cards, evaluation results, and governance documents described in the previous section. This category is different from the other four in that it is not evidence *of* something the AI produced; it is evidence *about* the system, and it is what makes categories (i) through (iv) provable or unprovable.

Four features make AI-touched evidence harder to authenticate than its predecessors, and they cut across the categories: content may be fabricated while appearing well-formed; provenance is often absent, because generated content carries no inherent origin marker; AI transformations may be embedded invisibly inside otherwise ordinary workflows; and, for synthetic media specifically, distinguishing generated from captured material is genuinely difficult even for qualified examiners.

For counsel, the operational value of this taxonomy is that it converts a vague objection into a specific request. If the concern is category (i), ask for interaction logs. If it is (ii), ask for the counterpart native file and its system metadata. If it is

(iii), ask for the original capture file, the device, and an unmodified copy with metadata intact. If it is (iv), ask for the independent records against which the disputed field can be checked. If it is (v), ask for the deployment record. A motion that asks for all of it, undifferentiated, invites a proportionality objection that will usually succeed.

KEY TAKEAWAYS

- “AI evidence” is not one category: generated content, AI-altered authentic content, wholly synthetic media, AI-suggested metadata, and the system's own operational records each have a different authentication path.
- The hardest category in practice is authentic material that has been partially altered, because the genuine portions corroborate the altered ones.
- AI-generated metadata is a real risk, but the mechanism matters: a model does not compute hash values — the exposure comes from human-entered or template-carried document properties that a forensic tool then reports faithfully.

The Authentication Baseline: How Rule 901 Already Works

Before examining where AI strains the rules, it is worth stating accurately what the rules already do, because a good deal of commentary in this area assumes a standard the Federal Rules do not impose.

Rule 901(a) requires the proponent of an item to produce evidence sufficient to support a finding that the item is what the proponent claims it is. That is all. It is a question of conditional relevance under Rule 104(b): the judge decides whether a reasonable juror *could* find the item genuine, and if so the item goes to the jury, which decides whether it *is*. Authenticity is ultimately a jury question. The judge is a screen, not a decider. Courts have observed for decades that the bar for authentication is not particularly high, and treating it as a demanding reliability inquiry is the most common analytical error in this area.

Rule 901(b) then supplies a non-exhaustive list of illustrations. Four carry most digital evidence.

901(b)(1) — testimony of a witness with knowledge. A participant to a conversation can authenticate a message thread; a custodian can authenticate a business system's output. This remains the workhorse.

901(b)(3) — comparison by an expert or the trier of fact against a specimen that has itself been authenticated. Relevant to voice and image comparison.

901(b)(4) — distinctive characteristics. The appearance, contents, substance, internal patterns, or other distinctive characteristics of an item, taken with all the circumstances. This is where digital evidence usually lives. A cryptographic hash computed at acquisition is a distinctive characteristic in the most literal sense available to the rule, and unaltered system metadata is another. A recognized practice is to designate an authenticating declarant who can speak with factual specificity about the process by which the electronically stored information was

created, acquired, maintained, and preserved without alteration — which is, in substance, a chain-of-custody declaration organized around 901(b)(4).

901(b)(9) — evidence about a process or system, describing the process or system and showing that it produces an accurate result. This is the provision that historically carried computer-generated output, and it is the provision the amendment proposals discussed in the next section would modify.

Rule 902 then supplies self-authentication routes that avoid live foundation testimony altogether. Rules 902(11) and (12) cover certified domestic and foreign business records. Rules 902(13) and (14), added in 2017, cover records generated by an electronic process or system and data copied from an electronic device or storage medium, in each case on a certification from a qualified person meeting the 902(11) or (12) requirements. These are underused. A properly drafted 902(14) certification describing the acquisition tool, the hash verification, and the certifier's qualifications will resolve most routine digital-evidence foundations without a witness, subject to the opponent's right to give reasonable written notice of an intent to challenge and state the particular grounds. What Rule 902 does not do is make evidence unassailable; it reallocates the effort.

Two further points of black-letter that matter here. Foundation evidence must itself be admissible — an unauthenticated screenshot cannot authenticate anything. And where processing for production may have altered the ESI, a proponent should be prepared to show that it did not, which is the practical basis of most form-of-production disputes. The most demanding published framework for electronic business records, articulated in *American Express Travel Related Services Co. v. Vinhnee (In re Vinhnee)*, sets out an eleven-step foundation running from the business's use of a computer through the reliability of its hardware and software to the identification of the exhibit; it is more than most courts require, and it is a useful checklist precisely because it is the ceiling rather than the floor.

Now the observation that organizes the rest of Part III, and it is more useful than most of what is written about deepfakes.

In 2011, courts confronted a new authentication problem: social-media evidence, and a defense that anyone could have accessed the account. A line of cases took that seriously enough to demand more than the ordinary showing — *Griffin v. State* in Maryland and *State v. Eleck* in Connecticut are the canonical skeptical decisions. A competing line applied the ordinary Rule 901 standard and asked whether a reasonable person could find the item genuine on the circumstantial proof available. The reasonable-person approach broadly prevailed, and it prevailed for a good reason: the skeptical approach made the mere possibility of fabrication a sufficient objection, which is not the standard the rule sets and would have made a large category of evidence practically unusable.

The 2026 deepfake objection is the same structure with better technology. It is a deniability claim: not that this exhibit was fabricated, but that it could have been, and therefore that the proponent should have to prove it was not. Courts already worked out that this is not what Rule 901 asks, and worked it out in a context where the risk of fabrication was real rather than hypothetical. That does not dispose of the problem — synthetic media is genuinely harder to detect than a spoofed login, and the next section takes the difference seriously. But it does mean the doctrine is not starting from nothing, and a court reaching for a framework has a well-tested one close at hand.

KEY TAKEAWAYS

- Rule 901(a) requires only evidence sufficient to support a finding that the item is what its proponent claims — a conditional-relevance question under Rule 104(b), and a deliberately low bar.
- Rules 901(b)(1), (b)(4), and (b)(9), together with the self-authentication provisions of Rule 902(11)–(14), already carry most digital evidence without any amendment.
- The 2026 “it's a deepfake” objection is structurally identical to the 2011 “my account was hacked” objection: a deniability claim that shifts the practical burden onto the proponent of genuine evidence.

Synthetic Media, the Liar's Dividend, and the Threshold Problem

Judges turn to this subject first, and usually with the wrong question. The pressing risk in litigation today is not that a court will unknowingly admit a fabricated video. It is that a party will neutralize authentic evidence by asserting that it might be fabricated, and that the assertion will cost the other side more to answer than it cost to make.

This asymmetry has a name in the academic literature: the **liar's dividend**. Once everyone knows convincing fakes exist, a party can attack authentic evidence simply by calling it a deepfake. The party making the claim bears almost no cost. The party defending genuine evidence may have to retain a forensic examiner, produce native files and device-level metadata, and litigate a collateral dispute — for an exhibit that was never in real doubt. Unmanaged, every digital exhibit becomes a potential mini-trial on authenticity, and the incentive structure rewards the party with the weaker case.

There are **two symmetric ways to get this wrong**, and a court that is guarding against only one is exposed to the other. The first is credulity: accepting a fabricated exhibit because it looks real and nobody had the resources to test it. The second is overcorrection: excluding or discounting genuine evidence because a party raised the possibility of fabrication and the court, uncertain, resolved the doubt against admission. The second error is currently the more common one, and it is invisible in the reported decisions, because a case that settles after a key exhibit is discounted leaves no record of why.

The threshold requirement is the answer to both. Before an authenticity inquiry opens, the objecting party should be required to state a specific and articulable basis to suspect fabrication — something about this exhibit, not something about the state of technology. “Deepfakes exist” is not a basis. “The audio contains a sentence the speaker was demonstrably elsewhere to have said, and the file bears

no capture metadata" is. Setting that threshold costs nothing and disposes of the tactical objection while leaving the genuine one fully available.

The most effective place to put it is in the **first case-management order, before any dispute has arisen**. A workable protocol has four parts, and none of them requires new rules:

1. **Provenance at source.** Where authenticity of audio, video, or images may be contested, require production of native files with original metadata intact, hash values computed at collection, and a documented chain of custody — rather than exported or re-encoded copies.
2. **A disclosure-and-challenge deadline.** Require any party intending to challenge the authenticity of an exhibit to identify the exhibit and state the basis by a date certain, well before trial or hearing.
3. **The threshold, stated in advance.** Recite the specific-and-articulable-basis standard in the order, so that it is a neutral rule of the case rather than a ruling against a party in the moment.
4. **A neutral examiner, reserved.** Provide that where a challenge meets the threshold, the court may appoint a neutral forensic examiner, with the allocation of cost addressed at that time.

Two structural points about proposals to change the rules. Serious amendment proposals have been put to the Advisory Committee on Evidence Rules. One line of proposal, associated with Judge Paul W. Grimm and Professor Maura R. Grossman, would tighten Rule 901(b)(9) — replacing the requirement that a process produce an accurate result with a requirement that it produce a valid and reliable result — and add a new provision addressing potentially fabricated or altered electronic evidence, with the court weighing probative value against the prejudicial effect of a fabrication. A second, associated with Professor Delfino, would shift the burden more decisively and would move the authenticity determination for digital audiovisual evidence from Rule 104(b) to Rule 104(a), taking it away from the jury. The two differ on the axis that matters most: whether

the fabrication question stays a conditional-relevance question for the jury or becomes a preliminary question for the judge. Commentators have opposed amendment on the ground that the existing rules already handle the problem and that a special rule risks implying that digital evidence is presumptively suspect. The Committee has to date preferred to wait for empirical evidence that the existing rules are failing, reasoning in part that countermeasures — watermarking and cryptographic fingerprinting at capture — may address the problem before a rule could take effect. **A separate proposal addressing machine-generated evidence, designated Rule 707, has been under consideration; readers should establish its text and current status from the Advisory Committee's own published agenda materials rather than from secondary descriptions, this one included.**

On the countermeasure the Committee pointed to: the **Coalition for Content Provenance and Authenticity (C2PA)** has published an open specification for cryptographically signed provenance metadata — commonly branded Content Credentials — which binds a signed statement about capture and subsequent edits to the media itself. Camera, phone, and software support has been expanding. Two limits should be understood before anyone treats this as a solution. A signature establishes what a signer asserted and that the file has not changed since; it does not establish that the depicted event occurred, and it says nothing at all about media that carries no credential, which is nearly all media. Whether a signed capture claim has been offered or accepted in a United States court is a question a reader should verify directly rather than assume; this guide does not assert an answer.

Finally, what courts have actually done so far is less dramatic than the commentary. The most-discussed early example is a California state-court products case against Tesla, in which counsel is reported to have suggested that recorded public statements attributed to the company's chief executive might be deepfakes, and the court to have been unpersuaded that the possibility excused compliance. A caution about this whole line of authority, and it is the reason no

case is cited for it here: these are trial-court orders, most of them unpublished and not carried in the public case-law databases, and they could not be verified against a primary source for this guide. A reader who intends to rely on one should obtain the order from the court's own docket rather than from any secondary description of it, including this one. What the reported orders appear to have in common is nonetheless instructive: courts facing a genuine dispute have responded by ordering production of the native file and its metadata — format, creation and modification dates, capturing device, lens, shutter speed — rather than by excluding the evidence or crafting new doctrine. That is the pattern to expect, and it is a sensible one: the dispute is resolved with discovery, not with a presumption.

KEY TAKEAWAYS

- The dominant practical risk is not that courts will admit fabricated video; it is that authentic evidence will be neutralized by an unsupported assertion that it might be fabricated.
- Requiring a specific and articulable basis before an authenticity inquiry opens — stated in a case-management order before any dispute arises — is the single most effective management tool available.
- Both symmetric failure modes are real: crediting a fabrication, and excluding genuine evidence out of an overcorrected fear of one.

Reliability, Expert Testimony, and the Rule 702 Gate

Rule 702 governs whether an expert may testify about any of this, and the analysis is not special to AI — which is the point. *Daubert v. Merrell Dow Pharmaceuticals, Inc.* identified a non-exclusive set of considerations for assessing the reliability of scientific methodology: whether the theory or technique can be and has been tested; whether it has been subjected to peer review and publication; the known or potential rate of error, and the existence and maintenance of standards controlling the technique's operation; and the degree of acceptance within the relevant community. *Kumho Tire Co. v. Carmichael* confirmed that the gatekeeping obligation extends to technical and other specialized testimony, not only to conventionally scientific evidence — which places forensic examination squarely inside it.

Applied to an AI-related expert, the factors sort themselves quickly.

Error rate is the pressure point, and it is not a formality. As Part I set out, there is no generally accepted error rate for synthetic-media detection, and detector performance degrades as generators improve. An expert offering a detection opinion should be asked, and should be able to answer: what tool, what version, what training or validation dataset, measured by whom, on what population, and when. If the answer is that the tool reports a confidence score, the follow-up is what that number means — a model's own confidence output is a property of the model, not a measured error rate, and the two are routinely conflated. Where the honest answer is that no validated error rate exists, that answer does not automatically exclude the testimony, but it should reshape it: the expert can competently describe what was observed and what it is consistent with, and cannot competently assign a probability to fabrication.

Peer review should be assessed by the reviewer, not by the fact of review. The question is the nature and standards of the reviewing body's procedures.

Publication in a venue with rigorous, disclosed review by a recognized standards or professional body is worth something; posting to a preprint server is not the same thing, whatever its merits for the field. And the ultimate allocation is unchanged: it is the court, not the scientific community, that determines whether a given opinion is reliable and relevant.

Validation chains matter more than they used to. Where one program relies on data generated by another piece of software, both must be validated — a principle that long predates AI and that AI-assisted workflows stress considerably. If a detection tool operates on frames extracted by a second tool from a container decoded by a third, an opinion about the first is only as good as the two beneath it. Counsel examining such an opinion should ask the expert to describe the full pipeline, and should expect a competent expert to have that answer ready.

The same gatekeeping instinct applies to the forensic report itself, and there is a compact framework for evaluating one that works for judges as well as for opposing counsel:

- **Reproducibility.** Can an independent examiner, given the same evidence, reach the same conclusion? If not, why not?
- **Acquisition mechanism.** Was a bit-for-bit forensic image taken, or was the evidence collected live from a running system? Both are defensible; they support different conclusions, and the report should say which was done and why.
- **Completeness of the package.** A report offered without the accompanying forensic images is a recognized warning sign, because it cannot be checked.
- **Tools stated.** The tools and versions used should be identified explicitly in the report. A reviewer needs to understand not just which tools were used but their provenance and known limitations — which maps directly onto asking about model provenance in an AI-assisted workflow.

- **Structure and omissions.** A report written to reach a conclusion rather than to report findings is a different document from a forensic report, and a material omission can be a serious matter rather than an oversight.
- **Acknowledged constraints.** Budget and time place real limits on any examination. A report that acknowledges what was not examined is more credible than one that implies exhaustiveness it did not achieve.

Two practical realities deserve mention because they shape outcomes more than doctrine does. First, the number of examiners who can do defensible work on synthetic media is small, and their engagement costs are substantial — sometimes tens of thousands of dollars for a single examination. That is an access-to-justice problem, not merely a budgeting one, and it is why the threshold requirement discussed above matters so much: a rule that lets any party trigger a forensic examination by assertion allocates a scarce and expensive resource to whoever is willing to be least scrupulous. Second, the technical expertise that qualifies a person to serve as a neutral in this area is frequently acquired through involvement with the industry or the technology at issue, which is exactly the involvement that must be disclosed. Courts and parties should expect meaningful disclosure statements from AI-qualified neutrals, and should not read a disclosure as disqualification.

KEY TAKEAWAYS

- The Daubert error-rate factor is the natural pressure point on a synthetic-media detection expert, because no generally accepted error rate exists for the task.
- Where one program consumes another's output, both must be validated — a chain that AI-assisted workflows extend rather than shorten.
- A forensic report that does not state the tools and versions used, and does not come with the underlying images, cannot be independently reproduced and should be weighted accordingly.

Discovery: Of AI Systems, and With AI Tools

Two problems are constantly conflated under one heading. Discovery **of** an AI system asks what a party's use of AI produced. Discovery **using** AI asks how the parties may use these tools to conduct review. They involve different rules, different objections, and different experts.

(A) Discovery of an AI system. The instinct is to demand the algorithm, and it is the wrong instinct in almost every case. Source code and model parameters are the most sensitive material a technology company holds, are protected as trade secrets, and — decisively — usually cannot answer the question anyway. A set of model parameters does not show what the system was asked, what it retrieved, or what it output on the day in question.

The proportionate targets, in rough order of value, are: the inference-time logs (prompts, tool calls, retrieved document identifiers, outputs, timestamps); the configuration in force at the relevant time, including model version, retention settings, and access controls; the deployment record (risk assessment, system or model card, evaluation and red-team results, governance minutes, accountable owner, incident-response plan); and the contract and data-processing terms with the provider, which establish what the provider holds and for how long. That list is specific enough to survive a proportionality objection under Rule 26(b)(1) and broad enough to reconstruct what happened.

Where the party did not hold the records — because the service is hosted — the practical route runs through the party's contractual right to obtain them before it runs through a subpoena to the provider, and the sequencing matters because retention windows are short. Where genuine trade-secret sensitivity attaches to something that must be examined, the established answer is a **secured-environment inspection**: the requesting party does not receive the material, a designated expert examines it in a controlled environment, may not remove anything from that environment, and reports findings to the court or tribunal under

a protective order. This mechanism is not new and is not AI-specific; it is the same accommodation courts have long made for source code, and it works here for the same reasons.

(B) Discovery using AI tools. Here there is a decade of directly applicable precedent that a surprising amount of current commentary ignores. Technology-assisted review — supervised classification applied to a document population — was approved for use in *Moore v. Publicis Groupe* and, by 2015, described in *Rio Tinto PLC v. Vale S.A.* as an accepted method where the producing party chooses to use it. The doctrine that developed around TAR is the doctrine that governs newer tools, and its core commitments are worth restating because they are exactly what is at risk of being forgotten:

- **Validation is not optional, and it is measured, not asserted.** A random sample, an agreed accuracy target, and an elusion test against the null set. A protocol without these is an assertion of completeness with no evidence behind it.
- **The training examples are part of the method.** Bias introduced through the seed set propagates. Introducing bias may increase precision while lowering recall, and worse, may cast doubt on the recall calculation itself.
- **Under-inclusivity is the standard attack.** A party using a predictive method opens itself to arguments about the adequacy of the data set searched and the selection criteria applied, and should expect them.
- **The producing party ordinarily chooses its method.** Courts have generally been reluctant to order a party to use a particular review technology, and a court mandating a specific method may fairly be said to be reaching beyond its role.

What does **not** yet exist is an equivalent settled protocol for generative-AI review — prompt-based relevance coding, retrieval-augmented search across a production, or LLM-assisted privilege review. The validation vocabulary transfers, but non-determinism means a sample validated on Monday does not straightforwardly

certify a run on Friday. Counsel proposing such a method should expect to specify the model and version, freeze it for the duration of the review, and validate with the same sampling discipline TAR requires; counsel opposing one should ask precisely those questions rather than objecting to the technology in the abstract. This is a genuine gap in the law, and pretending otherwise serves no one.

The ESI protocol is where all of this is actually decided, and it is decided early, quietly, and usually by default. A protocol negotiated the usual way is a set of defaults written by software, and software does not represent anyone's client. The clauses that matter most are unglamorous: whether search terms are applied with stemming, fuzzy matching, wildcards, or proximity, and what each of those settings does to the result; whether attachments and embedded objects are indexed at all, since a document that was collected but never searched is functionally not in the case; hit reports with unique-hit counts rather than raw totals; how modern attachments and hyperlinked cloud files are handled, given that a link is a pointer to a location and not a snapshot of content; the metadata field schedule; form of production; the scope of deduplication and what email threading discards; validation and sampling; time-zone normalization; and, for structured data, the data dictionary without which a database export is unreadable. Adding an AI clause to a protocol that gets those wrong is decoration.

One AI-specific protocol clause is worth negotiating explicitly: whether the parties must disclose the use of AI tools in creating, processing, or selecting the material they produce, and to whom. The emerging position in arbitration practice — that parties, and often the arbitrator, must be informed when AI tools are used in a matter — translates readily into a court protocol, and it is much easier to agree in advance than to litigate afterward.

KEY TAKEAWAYS

- Discovery of AI is almost never about “the algorithm”; the proportionate and answerable targets are logs, configuration, and the deployment record.

- Technology-assisted review is the established precedent for using a learning system in discovery, and its validation discipline — sampling, the null set, an elusion test — transfers directly to newer tools.
- There is as yet no settled validation protocol for generative-AI document review, which is a gap counsel should negotiate around explicitly rather than assume away.

Preservation and Spoliation of AI-Related Evidence

The duty to preserve attaches when a party reasonably anticipates litigation. That is settled and is not restated in the rule text; the 2015 rewrite of Federal Rule of Civil Procedure 37(e) deliberately left the trigger where the case law had put it.

What the 2015 amendment did settle is the consequence structure, and it made four choices that matter here. It confined the rule to electronically stored information. It rejected negligence — and gross negligence — as a sufficient predicate for the severest sanctions. It reserved the three severe measures (presuming the lost information was unfavorable, instructing the jury that it may or must so presume, or dismissing the action or entering default) to cases where the court finds the party **acted with the intent to deprive another party of the information's use in the litigation**. And for everything short of that, it requires a showing that the information should have been preserved, that it was lost because the party failed to take reasonable steps, that it cannot be restored or replaced through additional discovery, and that another party was prejudiced — with any measure no greater than necessary to cure the prejudice.

The distinction the rule draws is, in the useful formulation, between a fool and a knave. A new employee who wipes his desktop mail — including the litigation-hold notice he never read — has caused the same loss as a party who deleted the same mail to bury it. The rule treats them very differently, and correctly so. Nearly every contested Rule 37(e) motion is an argument about which one the court is looking at.

Applied to AI artifacts, four features make the analysis harder than it is for ordinary ESI.

Provider retention windows are short and contractual. Where an organization's email is retained for years by policy, its AI interaction logs may be retained for

days by default, or not at all under a zero-retention configuration. The duty to preserve does not extend the provider's retention clock. Counsel who anticipate that AI records will be relevant must act on that anticipation immediately — a hold notice issued three weeks into the matter may be issued into an empty bucket.

Session context is discarded by design. The material a model considered during a conversation typically exists as context for the duration of that session and is not itself stored unless the deployment logs it. What survives is what was logged, which is a configuration decision made long before the dispute.

The model can change beneath the evidence. A provider may update or replace the underlying model, alter its configuration, or change safety filters without notifying customers. Without a clear snapshot or audit trail, it can be effectively impossible to recover the state of an AI system as it existed at the time of a disputed event — and, unlike a deleted file, nothing was destroyed. The prior state simply is not retained anywhere. The closest familiar analogue is the hyperlinked-file problem: a link is a pointer to a location, not a snapshot of content, and an archiving tool exports the version that exists when the export runs, not the version that existed when the message was sent. Both problems have the same shape and the same partial solution, which is to capture contemporaneously or accept that you cannot.

Ephemeral channels are routinely under-collected, not through concealment but because standard collection workflows do not reach them. AI chat interfaces belong to that family, and a collection plan that covers mail, file shares, and a collaboration platform will silently miss them.

Two cautions for the party alleging spoliation. First, the principle from Part I applies with full force: **the absence of a record is not evidence that a record was destroyed.** On a great many deployments, no prompt log exists because logging was never enabled — a fact about procurement and configuration, not about intent. Establishing that logging *was* enabled, and that the logs are now missing, is a materially different showing from establishing that logs are absent, and the first

is usually available from the deployment record. Second, over-preservation is a real cost and courts are aware of it: the Advisory Committee heard directly from businesses that preserve mountains of data of which a tiny fraction ever proves relevant. A preservation demand for all AI-related data across an enterprise will draw a proportionality objection and will usually deserve one.

The practical drafting point: a litigation hold covering AI use should identify custodians, the specific systems and deployments, the date range, and the location of the records — including by department, IT system, provider, and, where relevant, geography. It should state expressly that provider-side retention settings must not be changed and that any scheduled deletion affecting the relevant systems must be suspended. And it should be issued to whoever controls the provider relationship, who is frequently not in the legal department and frequently has never seen a hold notice before.

KEY TAKEAWAYS

- Rule 37(e) reserves its severest sanctions for a finding that a party acted with the intent to deprive another party of the information's use — negligence, and even gross negligence, do not suffice.
- AI artifacts are unusually perishable: provider retention windows are short, session context is discarded, and the model itself can be replaced beneath the evidence.
- The absence of a prompt log is not proof that one was destroyed; on many deployments it is proof that logging was never enabled.

Questions Counsel Should Ask

The following converts Parts I through III into something usable at a Rule 26(f) conference, in a deposition, or in a first written request. None of it requires a technical background to ask; all of it requires one to answer.

Threshold questions, for every matter.

- Was any AI system used to create, alter, process, translate, summarize, or select any material relevant to this matter — by your client, by a vendor, or by a third party whose records are at issue?
- Which system, which version, and during what period? Is that version still available?
- Was the system hosted by the party or by a provider? Which provider, under what contract?

About the record.

- What is logged for each interaction — the prompt, the tool calls, the identifiers of retrieved documents, the output, the timestamp? Which of these are not logged, and why?
- What is the retention period for each category, and is it set by the provider's default, by the contract, or by the deployment's configuration?
- Is the deployment configured for zero retention? If so, since when, and who decided?
- Who holds the logs — the party or the provider — and what is the party's contractual right to obtain them?
- Has any retention setting been changed since the duty to preserve attached?

About the deployment.

- Is there a written AI risk assessment? A system or model card stating the intended purpose and known limitations?
- Was the deployment evaluated, tested, or red-teamed before it went into service? By whom, and what did the testing find?
- Who is the named accountable owner? What body approved the deployment, and are there minutes?
- What access controls govern what the system can retrieve? Have the system's tools or data sources been extended since deployment — and if so, was the permission record updated to reflect it?

About any expert opinion involving detection or classification.

- Which tool, and which version?
- What is its known or potential rate of error, measured by whom, on what dataset, and when?
- Is the reported number a measured error rate or the system's own confidence output? What is the difference in this instance?
- What was the full processing pipeline, and has every component in it been validated?
- Could an independent examiner, given the same inputs, reproduce this result? If not, precisely what would differ, and why?
- When was this tool last validated against content generated by current systems?

About the exhibit itself, where authenticity is contested.

- Is the native file available with its original metadata intact, or only an export or re-encode?
- Was a hash computed at collection? By whom, using what tool, and has it been re-verified since?

- What is the documented chain of custody from capture to production?
- What independent records corroborate the disputed field — mail system logs, document management history, device records, testimony from a participant?
- What is the specific and articulable basis for suspecting fabrication, as distinct from the general availability of the technology?

A note on sequencing. Almost all of these questions are cheaper to ask in the first thirty days than in the ninth month, and several of them — the retention questions in particular — become unanswerable if they are asked late. The single highest-value change most litigation teams can make in this area is to move the threshold question to the top of the initial case-assessment checklist and stop treating it as a specialty topic.

KEY TAKEAWAYS

- The threshold question in every matter is now simply whether AI was used to create, alter, process, or select any evidence — and it should be asked at the Rule 26(f) conference, not discovered at deposition.
- For any detection or classification opinion, the four questions that matter are: which tool, which version, validated against what, and is the result reproducible by an independent examiner from the same inputs.

Questions the Court Will Face, and the Tools Available to It

Four decisions land on the bench in AI-related matters, and they are worth naming separately because they are often argued as one.

Does an authenticity inquiry open at all? This is the threshold question from Part III, and it is the most consequential because it is the one that determines cost. A court that requires a specific and articulable basis before opening an inquiry has resolved most tactical objections without reaching the merits of any. A court that treats the general availability of synthetic media as a sufficient basis has effectively made every digital exhibit contestable.

Who decides authenticity — judge or jury? Under current Rule 901 and Rule 104(b), the judge screens and the jury decides. One of the pending amendment proposals would move digital audiovisual evidence to Rule 104(a), making the judge the decider. A court applying the rules as they stand should be clear about which allocation it is using, because the difference is not procedural nicety: it determines whether the jury hears the evidence at all.

How is the cost allocated? Forensic examination in this area is expensive and the expertise is scarce. Under the American Rule, expert fees are generally treated differently from attorney's fees, and a party with resources can impose examination costs on a party without them simply by raising a challenge. Proposals to shift costs where a challenge is unsupported, screened by an assessment of the parties' relative means and by Rule 11-type sanctions against tactical allegations, are under discussion. A court need not wait for a rule: the threshold requirement, plus a stated willingness to allocate the cost of an unsupported challenge to the party who raised it, addresses most of the problem at the first conference.

How does the court avoid both errors? Crediting a fabrication and excluding genuine evidence out of an overcorrected fear of one are both real failure modes, and it is uncomfortable but accurate to observe that outcomes in this area currently turn largely on an individual decision-maker's comfort with the technology. That is an argument for structured procedure, not for greater confidence. A decision-maker who requires a threshold showing, orders provenance production, and appoints a neutral where a challenge is genuinely supported does not need to be a technologist to reach a defensible result.

The **toolkit already exists**, and none of it requires amendment:

- **Rule 16 conferences.** The single most effective intervention available. A paragraph in the first procedural order — requiring disclosure of AI use, setting the challenge deadline, stating the threshold, and directing native production with metadata for contested media — does work that the rulemaking process, measured in years, cannot do in time. This is one respect in which arbitration and case-managed litigation have a structural advantage over rulemaking: a tribunal can adopt a workable protocol immediately.
- **Rule 26 expert disclosures.** The report requirement, used rigorously, surfaces most of what Part III's Rule 702 questions are looking for before any hearing.
- **Federal Rule of Evidence 706 — court-appointed experts.** Available, underused, and directly suited to a dispute in which both parties' experts are advocating and the court needs a technical answer rather than a contest. Judges have historically been reticent to appoint, partly from concern about how the appointee is selected; a workable method is to have each side's expert participate in choosing a third.
- **Federal Rule of Civil Procedure 53 — special masters.** Well established for e-discovery supervision, and a natural fit for AI-system inspection under a protective order, for supervising a secured-environment examination, and for managing validation disputes. A forensically qualified neutral can determine the existence and authenticity of digital evidence in a way that adversarial

expert exchange often cannot, and in some circumstances — a party ordered to search systems the other side cannot see — a neutral technologist is the only effective means of confirming compliance.

- **Rule 1006 summaries and demonstratives.** Where the underlying material is voluminous logs, a summary exhibit is usually the only realistic path to a comprehensible presentation.
- **Concurrent expert evidence.** The practice sometimes called hot-tubbing, in which opposing experts testify together and address each other's positions in front of the decision-maker, is well suited to technical disputes in which the disagreement is narrow and buried.

A closing observation on judicial education. Several proposals in this area converge on the same conclusion from different directions: that baseline technical training for judges and attorneys would do more good than a new rule, and that a continuing-education requirement modeled on the technology and cybersecurity requirements some jurisdictions already impose is a plausible vehicle. That is not a substitute for the procedural tools above. It is what makes them usable.

KEY TAKEAWAYS

- A paragraph in a procedural order entered at the first conference can do work that would take the federal rulemaking process years to accomplish.
- The court already has a full toolkit — Rule 16 conferences, Rule 26 expert disclosures, FRE 706 court-appointed experts, Rule 53 special masters, Rule 1006 summaries, and concurrent expert evidence — and none of it requires amendment to reach AI.
- Outcomes in this area currently turn too much on an individual decision-maker's comfort with the technology, which is an argument for structured procedure rather than for greater confidence.

Professional Responsibility, Governance, and the Rules That Already Exist

The professional-responsibility dimension is the part of this subject every reader has heard about and the part most often described badly. The organizing principle is simple and worth stating before any of the detail: **AI does not alter a lawyer's preexisting professional duties.** It supplies a new instrument to which existing duties attach without modification.

The competence duty already reaches it. Comment 8 to ABA Model Rule 1.1 provides that to maintain the requisite knowledge and skill, a lawyer should keep abreast of changes in the law and its practice, including the benefits and risks associated with relevant technology. That language was written for a different generation of tools and applies to this one without strain.

The sanction cases are about verification, not about use. In *Mata v. Avianca, Inc.*, counsel submitted a brief containing citations to judicial decisions that did not exist, produced by a generative tool, and then did not withdraw them when their existence was questioned. The court's stated basis for sanctions was not that the lawyers used AI; it was that they abandoned their responsibilities when they failed to verify its output and then persisted. In *Park v. Kim*, the Second Circuit confronted a brief citing a nonexistent decision generated by the same class of tool and referred the attorney to its grievance panel. The pattern in both, and in the growing body of similar orders, is consistent: the breach is of the duty to verify and of the duty of candor to the tribunal. Rule 11(b) already reaches it, which is one of the arguments made against adopting special AI certification requirements.

Disclosure requirements are proliferating and are not uniform. Individual federal judges have adopted standing orders of several different shapes: mandatory certification that either no generative AI was used in drafting or that any AI-generated language was checked by a human; requirements to identify the specific program and the portions of the filing it produced, coupled with a

certification about confidentiality; and, in at least one instance, a condition on admission. At least one federal court of appeals has considered and declined to adopt a circuit-wide rule, on the ground that Rule 11(b) already covers the conduct. **These orders change frequently.** The only reliable practice is to check the individual judge's standing orders and the local rules at the outset of each matter, every time. Any list of them, including any list a reader might find in secondary literature, should be assumed stale.

State bar guidance has converged on a common core. A substantial number of state bars and bar task forces have issued opinions or reports on generative AI in practice since late 2023, and while the details differ, the through-line is consistent: limited-use transparency, disclosure where material, human oversight, verification of output, protection of client confidences, and care in billing for time not actually spent. At least one such opinion is addressed specifically to judicial officers rather than to practitioners. Because this landscape moves quickly and by jurisdiction, this guide deliberately does not reproduce a list; a practitioner should consult their own jurisdiction's current guidance directly.

Confidentiality deserves specific attention because it is the duty most easily breached without noticing. Submitting client material to a third-party service is a disclosure to that service. Whether it is a permissible one depends on the terms — whether the provider trains on customer content, what it retains, for how long, who can access it, and what happens on subpoena. Consumer-tier and enterprise-tier terms for the same product frequently differ on exactly these points. A firm that has not read its provider's terms has not made a considered decision about client confidences.

A verification protocol is straightforward to implement and is the single most effective control available. Three elements suffice for most practices: restrict AI use to approved, sandboxed systems with known retention terms; require that every authority produced with AI assistance be opened and read in a primary source before it appears in a filing, recorded in a signed source-verification log;

and run a final check of the table of authorities against a citator before filing. None of this is exotic, and all of it is cheaper than one show-cause hearing.

Governance frameworks matter to litigators for two reasons, neither of which is compliance for its own sake. They increasingly define the standard of care, and they generate the documentary record that Part II identified as the actual discoverable material. Three are worth knowing by name. The **NIST AI Risk Management Framework** (NIST AI 100-1) organizes AI risk management around four functions — GOVERN, MAP, MEASURE, and MANAGE — and is expressly voluntary and adaptable rather than prescriptive; its accompanying playbook offers suggested actions, not requirements. **ISO/IEC 42001:2023** specifies requirements for an AI management system and is certifiable, which makes the presence or absence of certification a discoverable fact. And the joint guidelines on secure AI system development published by the United States Cybersecurity and Infrastructure Security Agency together with the United Kingdom's National Cyber Security Centre address the development lifecycle from a security perspective. Outside the United States, the **EU AI Act** (Regulation (EU) 2024/1689) imposes record-keeping and automatic event-logging obligations on high-risk systems; its obligations phase in on a staged timetable, so any statement about what it currently requires, of whom, needs a date check against the regulation itself.

Two framing points close this out. First, explainability and traceability are not abstractions in this context; they are the prerequisites for forensic readiness. An organization that cannot reconstruct what its system did cannot defend what its system did. Second, and against a persistent misconception: **automation concentrates responsibility within an organization's processes; it does not eliminate it.** Where a legal standard turns on what humans knew or had reason to know, the deployment record is the evidence of what they knew, and an organization that chose not to create one has not thereby made the question go away.

KEY TAKEAWAYS

- Sanctions in this area have followed the failure to verify, not the use of the tool — the duty breached is diligence and candor, not a prohibition on technology.
- AI does not alter a lawyer's preexisting professional duties; competence, confidentiality, candor, and supervision apply unchanged to a new instrument.
- Governance frameworks matter to litigators for two reasons: they increasingly define the standard of care, and they generate the documentary record that Part II identifies as the discoverable material.

A Short Glossary

Agent (agentic system). A generative model configured to pursue a goal by calling tools, querying data sources, and executing multi-step tasks with limited human supervision.

Algorithm. A specific sequence of steps that accomplishes a defined operation. Not a synonym for model, tool, or system.

Chain of custody. The documented sequence of persons and processes that held or handled an item of evidence from collection to presentation, together with the integrity checks applied at each transfer.

Context window. The maximum amount of material a model can consider in a single interaction. Anything outside it was not before the system.

Deepfake. Audio, video, or imagery that has been generated or substantially altered by machine learning so as to depict a person doing or saying something they did not.

Elusion test. A validation method in which a random sample is drawn from the material a search or classifier rejected, reviewed by a human, and used to estimate how much responsive material the process missed.

Embedding. A numeric representation of text or an image that places semantically similar items close together, enabling retrieval by meaning rather than by exact term.

Fine-tuning. Additional training applied to an existing general-purpose model to specialize it for a task or domain.

Forensic image. A bit-for-bit copy of a storage medium, including deleted and unallocated space, made through a write blocker and verified with a cryptographic hash.

Hallucination. Output that is fluent and well-formed but factually false, produced because the system predicts likely continuations rather than verifying claims. Not a malfunction.

Hash (cryptographic digest). A fixed-length value computed from a file's contents by a one-way function, such that any change to the contents produces a completely different value. Used to prove that two copies are identical and that a copy has not changed.

Inference. The act of running a trained model on an input to produce an output, as distinct from training.

Large language model (LLM). A generative model trained on very large volumes of text to predict likely continuations, and the basis of current chat assistants.

Model. The specific trained artifact — in current systems, a large set of numeric parameters produced by training. Distinct from the tool that packages it and from the deployment that configures it.

Model card / system card. A document published by a developer or deployer describing a system's intended purpose, training approach at a general level, evaluation results, and known limitations.

Null set. The material a search or classification process did not return. Where an unvalidated process hides its failures.

Precision. The share of the items a process returned that are genuinely responsive.

Prompt. The input submitted to a generative system, including any instructions, context, and retrieved material placed before the model.

Provenance (content). Verifiable information about where a piece of content came from and how it was modified. Content Credentials, under the C2PA specification,

are the leading current mechanism for attaching signed provenance to media.

Recall. The share of all genuinely responsive items that a process actually found.

Retrieval-augmented generation (RAG). A pattern in which a system searches a document collection and places what it finds into the model's context before generating an answer.

Richness. The proportion of a document population that is genuinely responsive. A property of the collection, not of the tool.

Slack space / unallocated space. Storage that the file system regards as available for reuse but which may still contain the contents of deleted files, until reclaimed.

Technology-assisted review (TAR). Supervised classification applied to a document population in discovery, trained on human coding decisions and validated by sampling.

Token. The unit a language model reads and writes — roughly a word fragment.

Training. The process that produces a model from data, distinct from and usually far more expensive than inference.

Write blocker. A hardware or software device placed between source media and an acquisition system so that the acquisition process cannot alter the source.

KEY TAKEAWAYS

- Definitions here are deliberately short and non-technical; each is written to be usable in a brief or from the bench without further translation.

Method, and What This Guide Does Not Claim

This guide was assembled from two kinds of material. The first is a body of first-party writing on digital evidence, forensic examination, e-discovery, authentication, arbitration practice, and AI governance, produced over roughly fifteen years and used here as a finding aid — a way of locating the underlying rules, cases, and problems that practitioners actually encounter. The second is primary authority: the Federal Rules of Evidence and Civil Procedure, decided cases, and published standards, each consulted directly.

The two halves of this guide do not rest on the same material, and a reader should weigh them accordingly. Parts III and IV — authentication, expert reliability, discovery, preservation, and the case-management toolkit — draw substantially on that first-party body of work, which is deep in precisely those areas and was written for judges and litigators. Parts I and II — what these systems are, how they behave, and what records they leave — do not, because that same body of work is thin on the mechanics: it is writing by lawyers about evidence, not by engineers about models. Terms such as transformer, embedding, context window, and retrieval-augmented generation barely appear in it. Those sections were therefore built from primary technical, standards, and rulemaking material, and every mechanical description in them was written to be defensible on cross-examination rather than to sound authoritative. Where the honest answer was that no verifiable source could be established — an error rate for a detection product, a count of judicial standing orders, the disposition of a particular unpublished trial-court order — the guide says so in place of the figure.

Three commitments shaped what appears above, and it is worth stating them because they explain some of the omissions.

Nothing is cited that was not verified against a primary source. Case citations in the Sources list below were checked against a public case-law database; where a reported citation could not be confirmed there, the docket was confirmed instead

and the citation is given accordingly. Where a proposition was supported only by a secondary description, the guide states the substance and directs the reader to the primary source rather than supplying a citation that would look more authoritative than the underlying evidence warrants. A named gap is a professional disclosure; a citation that cannot be checked is not.

Several widely circulated figures are deliberately absent. These include: point-in-time counts of how many federal judges have issued AI standing orders; aggregate loss figures attributed to deepfake fraud; percentages of practitioners or organizations using AI tools; and stated error rates for synthetic-media detection products. Each of these circulates in the practitioner literature, in some cases with a citation to another piece of practitioner literature. None is repeated here, because no verifiable primary source for it was established. Readers who need such a figure should obtain it from the body that produced it and should note its date, because every one of them is a snapshot.

Rights in the underlying material were respected. A substantial portion of the first-party writing consulted has been published by third parties — legal publishers, bar and dispute-resolution institutions, and the federal judiciary — under licences that reserve rights in the language. Where a source is encumbered in that way, it was used to locate the underlying authority and the prose here was written fresh. The Sources list records the rights status of each item so that a reader or an institution can see what stands behind a given proposition and on what terms. Items marked as rights unconfirmed are exactly that: unresolved, not free.

On the use of AI in preparing this guide. A document that asks others to disclose their use of these tools should account for its own. This guide was prepared with machine assistance in searching and organizing a large body of source material. Every legal authority cited was verified against a primary source; every technical description was written to be defensible on cross-examination rather than merely fluent; and no citation, quotation, statistic, date, or authority appears here that was

not confirmed. Where confirmation was not possible, the point is stated without a citation and flagged for the reader to verify, and those flags are deliberate.

Corrections. This guide will be updated. Readers who identify an error — particularly in a citation, a rule reference, or a technical description — are encouraged to write to info@lawandforensics.com so that the correction can be made in the next revision and, where the error is material, noted in the revision history.

KEY TAKEAWAYS

- Every case, rule, and standard cited here was checked against a primary source; where a proposition could not be verified to that standard, the guide states the substance and tells the reader to verify rather than supplying a citation.
- The evidence and procedure sections rest largely on a first-party body of legal writing; the sections describing how the technology works do not, and were built from primary technical and standards material instead.
- Several propositions that circulate widely in this area — point-in-time counts of judicial standing orders, loss figures for deepfake fraud, and error rates for detection tools — are deliberately omitted because no verifiable primary source was established for them.

Frequently asked questions

Does a party have to disclose that it used AI to draft a filing?

It depends entirely on the forum. There is no uniform federal requirement. A number of individual district judges have adopted standing orders requiring certification, disclosure of the specific program used, or identification of the portions of a filing produced by AI, and at least one court of appeals has considered and declined to adopt a circuit-wide rule on the ground that Rule 11(b) already reaches the conduct. Several state bars have issued guidance. Because these orders change frequently and are judge-specific, the only reliable practice is to check the assigned judge's standing orders and the local rules at the outset of every matter.

If a party claims a video exhibit is a deepfake, who has to prove what?

Under the Federal Rules as they currently stand, the proponent must produce evidence sufficient to support a finding that the item is what it is claimed to be — the ordinary Rule 901(a) standard, treated as a question of conditional relevance under Rule 104(b). The objecting party does not carry a formal burden, but courts increasingly require a specific and articulable basis for suspecting fabrication before opening a forensic inquiry, because the general availability of synthetic media would otherwise make every digital exhibit contestable at no cost to the challenger. Pending amendment proposals would change this allocation in different ways, including one that would move the determination for digital audiovisual evidence to Rule 104(a) and take it from the jury.

What should I actually ask for in discovery when the other side used an AI system?

Not the algorithm. Ask for the inference-time logs — the prompts submitted, the tools the system called, the identifiers of documents it retrieved, the outputs returned, and the timestamps; the configuration in force at the relevant time, including model version and retention settings; the deployment record, meaning the AI risk assessment, the system or model card, evaluation and red-team results, governance minutes, and the named

accountable owner; and the contract with the provider, which establishes what the provider holds and for how long. That set is specific enough to survive a proportionality objection and sufficient to reconstruct what happened.

Can an AI system's output be authenticated under Rule 901(b)(9) as the product of a process or system?

Rule 901(b)(9) requires a description of the process or system and a showing that it produces an accurate result, and that second element is where generative systems create difficulty: their outputs are probabilistic, not deterministic, and the same input can produce different outputs. What can usually be authenticated is the record of the interaction — that this system, in this version and configuration, was given this input on this date and returned this output — which is a materially different exhibit from the assertion contained in the output. One pending amendment proposal would replace the rule's "accurate result" language with "valid and reliable result" partly to address this.

How long are AI prompt logs kept?

Often not long, and it varies by provider, by product tier, and by how the customer configured the deployment. Default retention for enterprise AI services is commonly measured in days or weeks rather than years, and some enterprise configurations retain nothing at all after a session ends. The duty to preserve does not extend a provider's retention clock. If AI records may be relevant, the retention terms and the deployment's configuration should be established in writing in the first days of the matter, and the hold notice should reach whoever controls the provider relationship.

Is a missing prompt log evidence of spoliation?

Usually not, on its own. On a great many deployments no prompt log exists because logging was never enabled — a procurement and configuration fact, not a fact about anyone's intent. Establishing that logging was enabled and that the logs are now missing is a materially different showing from establishing that logs are absent, and the first is normally provable from the deployment record. Rule 37(e) reserves its severest measures for a finding that a

party acted with the intent to deprive another party of the information's use, and rejected negligence and even gross negligence as sufficient predicates.

Are Content Credentials or C2PA signatures a solution to synthetic media in court?

They are a genuine improvement and not a solution. A cryptographic provenance signature establishes what a signer asserted about capture and editing, and that the file has not changed since it was signed. It does not establish that the depicted event occurred, and it says nothing about media that carries no credential — which is nearly all media currently in circulation. Whether a signed capture claim has been offered or accepted in a United States court is a question a reader should verify directly rather than assume.

Does using technology-assisted review or AI-assisted review create a risk of sanctions?

Using it does not; using it without validation might. The established discipline from a decade of technology-assisted-review practice is a random sample, an agreed accuracy target, and an elusion test drawn from the material the process rejected. A methodology with no validation is an assertion of completeness with no evidence behind it, and that is the exposure. For generative-AI review specifically there is as yet no settled validation protocol, so a party proposing one should expect to specify and freeze the model version and validate with the same sampling discipline.

Sources

Fed. R. Evid. 901 (Authenticating or Identifying Evidence) — 901(a); 901(b)(1), (3), (4), (9)

Fed. R. Evid. 902 (Evidence That Is Self-Authenticating) — 902(11)-(14)

Fed. R. Evid. 104 (Preliminary Questions) — 104(a); 104(b)

Fed. R. Evid. 702 (Testimony by Expert Witnesses)

Fed. R. Evid. 706 (Court-Appointed Expert Witnesses)

Fed. R. Evid. 1006 (Summaries to Prove Content)

Fed. R. Civ. P. 37(e) (Failure to Preserve Electronically Stored Information) — 37(e)(1)-(2)

Fed. R. Civ. P. 26 (Duty to Disclose; General Provisions Governing Discovery) — 26(b)(1) proportionality; 26(f) conference

Fed. R. Civ. P. 53 (Masters)

Fed. R. Civ. P. 16 (Pretrial Conferences; Scheduling; Management)

Fed. R. Civ. P. 11(b) (Representations to the Court)

Advisory Committee on Evidence Rules, agenda books and reports (Judicial Conference of the United States) — the primary source for the status of proposed amendments to Rule 901 and of proposed Rule 707

Daubert v. Merrell Dow Pharmaceuticals, Inc., 509 U.S. 579 (1993) — 509 U.S. at 593-94 (reliability factors)

Kumho Tire Co. v. Carmichael, 526 U.S. 137 (1999) — gatekeeping extends to technical and specialized testimony

Lorraine v. Markel American Insurance Co., 241 F.R.D. 534 (D. Md. 2007)

American Express Travel Related Services Co. v. Vinhnee (In re Vinhnee), 336 B.R. 437 (B.A.P. 9th Cir. 2005) — eleven-step foundation for electronic business records

Griffin v. State, 419 Md. 343, 19 A.3d 415 (2011)

State v. Eleck, 130 Conn. App. 632, 23 A.3d 818 (2011)

Moore v. Publicis Groupe, 287 F.R.D. 182 (S.D.N.Y. 2012) (commonly cited as Da Silva Moore)

Rio Tinto PLC v. Vale S.A., 306 F.R.D. 125 (S.D.N.Y. 2015)

Mata v. Avianca, Inc., No. 1:22-cv-01461 (S.D.N.Y. June 22, 2023) (Castel, J.), Opinion and Order on Sanctions, ECF No. 54 — reported at 678 F. Supp. 3d 443; that reporter citation did not resolve in the public case-law database consulted, so the docket entry is given as the source of record — ECF No. 54, at 1 ("Technological advances are commonplace and there is nothing inherently improper about using a reliable artificial intelligence tool for assistance. But existing rules impose a gatekeeping role on attorneys to ensure the accuracy of their filings.")

Park v. Kim, 91 F.4th 610 (2d Cir. 2024) — referral of counsel to the court's Grievance Panel under Local Rule 46.2, counsel having cited a non-existent state-court decision identified by a generative AI tool without confirming it

Zubulake v. UBS Warburg LLC, 217 F.R.D. 309 (S.D.N.Y. 2003)

NIST, Artificial Intelligence Risk Management Framework (AI RMF 1.0), NIST AI 100-1 (January 2023) — the GOVERN, MAP, MEASURE, MANAGE functions

NIST Trustworthy and Responsible AI Resource Center — current NIST work on synthetic content, provenance, and evaluation

NIST Special Publication 800-86, Guide to Integrating Forensic Techniques into Incident Response (2006)

ISO/IEC 42001:2023, Information technology — Artificial intelligence — Management system

Coalition for Content Provenance and Authenticity (C2PA), C2PA Specification (Content Credentials)

Regulation (EU) 2024/1689 (EU Artificial Intelligence Act) — record-keeping and automatic event-logging obligations for high-risk systems; obligations phase in on a staged timetable — Art. 12 (record-keeping)

Cybersecurity and Infrastructure Security Agency and UK National Cyber Security Centre, Guidelines for Secure AI System Development

ABA Model Rules of Professional Conduct, Rule 1.1 (Competence), cmt. 8 — cmt. 8

Law & Forensics working paper, Key Considerations for Authenticating Evidence Under Rule of Evidence 901 As We Enter the Era of Generative AI (2025, unpublished) — doc#956, headings “Introduction > What is Generative AI?”; “Authenticating Audio/Visual Evidence Under Rule 901”; “Proposals to Amend Rule of Evidence 901”; chunks 4-12, 13-30, 39-48

Daniel B. Garrie and Morgan B. Ward Doran, Agentic AI and the Limits of Scienter Under the Federal Securities Laws (draft) — doc#1002 ¶¶0, 8, 9, 11 (logging, deployment record, incident-response preservation)

Daniel B. Garrie and Hon. Leo M. Gordon, 10 ESI Protocol Clauses That Decide Cases Before the First Document Is Read (2026) — doc#990, headings “Search Terms, Fuzzy Matching, and Proximity”; “Hit Reports”; “Modern Attachments and Hyperlinked Files”; “Validation and Sampling”

Daniel B. Garrie, Stacy Harrison and Ryan Booms, Digital Forensic Investigations and eDiscovery, ch. 6 of Plugged In: Guidebook to Software and the Law (2023 ed.) — doc#1935 ¶¶3-13 (file systems, deletion, overwriting, metadata), ¶¶17-21 (report-evaluation framework), ¶¶24-28 (supervised review and its metrics)

Hon. James Ware, Hon. Margaret M. Sweeney and Daniel B. Garrie, A Procedural Framework for Educating Decisionmakers About Software, ch. 9 of Plugged In: Guidebook to Software and the Law (2023 ed.) — doc#1921 ¶¶2, 10, 29-37, 44, 48 (Rule 16, Rule 706, Rule 53, Rule 1006, concurrent expert evidence)

Daniel B. Garrie and Hon. John M. Facciola, E-Discovery and the Federal Rules of Civil Procedure Relating to Electronic Data, ch. 10 of Plugged In: Guidebook to Software and the Law (2023 ed.) — doc#1922 ¶¶31-35 (the 2015 Rule 37(e) rewrite), ¶33 (over-preservation), ¶43 (the fool/knave illustration)

eDiscovery for Corporate Counsel, ch. 6 (2017/18 ed.) — doc#1448 ¶¶25-26 (hash values as Rule 901(b)(4) distinctive characteristics; the authenticating declarant)

eDiscovery for Corporate Counsel, ch. 16 (2017/18 ed.) — doc#1458 ¶¶15, 17, 23, 42-43 (FRE 901-903, 902(7)/(11)/(12), 1002; In re Vinhnee)

Law & Forensics working paper, Seeing Is No Longer Believing: How Arbitrators Authenticate in the Deepfake Era (2026) — byline unresolved in the source — doc#996, headings “How the Dividend Is Earned”; “Two Ways to Get It Wrong”; “The Threshold Requirement”; “A Recommended Protocol”

Daniel B. Garrie and Hon. Leo M. Gordon, The Hallucination Tax: Why AI Negligence Is a Self-Inflicted Wound on Arbitral Efficiency (2026) — doc#995 ¶¶0-1 (hallucination as a structural property), ¶¶7-8 (three-phase verification protocol)

Daniel B. Garrie and David Cass, Governing Artificial Intelligence as Enterprise Infrastructure: A CISO Perspective (2026) — doc#991 chunk [2] (explainability and traceability as prerequisites for forensic readiness); NIST AI RMF and ISO/IEC 42001 framing

Daniel B. Garrie and Jennifer Deutsch, Detecting Deception: Tools for Identifying Deep Fakes in Court Proceedings (2024 draft) — doc#909 ¶¶1-3 (three detection families; absence of consensus on an acceptable error rate)

Daniel B. Garrie, Jennifer Deutsch and Morgan B. Ward Doran, AI-Generated Metadata Hallucination: A New Risk for Attorneys and the Legal Community (Law360 2025) — doc#954 ¶¶6-7 — used to locate the underlying risk only; an executed publisher licence exists at doc#955

Daniel B. Garrie and Hon. Charles Margines (Ret.), Deciphering Digital Dilemmas: Forensic Neutrals in Large-Scale Litigation (2025) — doc#960 ¶¶1-4 (grounds for appointing a neutral; functions of a forensic neutral)

Daniel B. Garrie and James Mariani, Authenticating Social Media Evidence (Thomson Reuters Legal Executive Institute, 2018) — doc#534 pp. 2-5 — the skeptical and reasonable-person authentication lines; used to locate the cases only

Daniel B. Garrie, Hon. Leo M. Gordon and Bradford K. Newman, Uncovering Digital Evidence: A Comprehensive Guide for Legal Professionals in the Digital Era (Springer, 2024) — doc#947 pp. 164, 171, 179 — exclusively licensed to the publisher; consulted for structure and to locate authorities, not reproduced

Understanding Software (Federal Judicial Center, 2015), based on Garrie & Allegra, Plugged In (Thomson Reuters 2013) — doc#1136 pp. 6, 43-56 — distribution restricted to the federal judiciary; used as a genre model only, with no language reproduced

Daniel B. Garrie and Jeremy Desor, The Importance of Curiosity in Conducting a Complex White Collar Fraud Investigation (2026) — doc#999, heading “Resisting the First Theory” (routine under-collection of ephemeral channels)

Related reading

INSIGHT

Authenticating AI-Generated Evidence in the Courtroom

INSIGHT

AI-Generated Metadata Hallucinations: A New Risk for Attorneys

INSIGHT

AI Without Guardrails Is a Liability: Building the Governance Framework Your Organization Actually Needs

INSIGHT

Digital Forensic Evidence in the Courtroom: Understanding Content and Quality

IN THIS GUIDE

Method, and What This Guide Does Not Claim