

THE FEDERAL COURTS LAW REVIEW‡

Volume 18

2026

BEYOND TAR: VALIDATING HUMAN AI-ASSISTED REVIEW IN eDISCOVERY

*Neil Cameron**

INTRODUCTION: THE VALIDATION GAP.....	2
I. THE DOCTRINAL ARCHITECTURE OF TAR	
VALIDATION	7
A. <i>The Epistemic Foundations of Recall and Precision</i>	7
B. <i>The Sedona Framework and the Transparency Norm</i>	9
C. <i>Rule 26(g) and the “Reasonable Inquiry” Standard</i>	10
II. THREE POSITIONS IN THE CURRENT DEBATE	11
A. <i>The Continuity Position</i>	12
B. <i>The Reliance-Point Model</i>	14

‡ The *Federal Courts Law Review* is a publication of the Federal Magistrate Judges Association. Editing support is provided by the members of the *Mississippi Law Journal*. Note on Sources: This article draws on original interviews conducted by the author in 2026 with Judge Paul W. Grimm (ret.), Professor Maura R. Grossman, Professor Jason R. Baron, and Jim Sullivan (CEO, eDiscovery AD). All interview quotations are used with the interviewees’ consent and have been verified for accuracy. Additional interviews were conducted with David Stanton (Partner, Pillsbury Winthrop Shaw Pittman LLP) and with practitioners at Consilio, Epiq, OpenText, Reveal, Control Risks, Everlaw, and QuisLex. The author is grateful to all participants for their generosity.

* Neil Cameron, Lead Analyst at Legal IT Insider and Affiliate Scholar at the Judith Reppy Institute for Peace and Conflict Studies at Cornell University.

<i>C. The Discontinuity Position and the Medicine</i>	
<i>Analogy</i>	18
III. THE OBJECT-OF-PROOF PROBLEM AND THE	
NINE-LAYER FRAMEWORK	19
<i>A. The Causation Proposition</i>	19
<i>B. A Nine-Layer Validation Architecture</i>	21
Layer 1: Function Mapping and Use Case	
Categorization	21
Layer 2: Output Type Characterization	22
Layer 3: Statistical Validation for Binary Outputs	22
Layer 4: Source Document Integrity Verification.....	23
Layer 5: Contextual Coherence Sampling	23
Layer 6: Pre-Deployment Independent Validation	23
Layer 7: Source-Anchored Interpretive Verification.....	24
Layer 8: Structured Human Override Protocols	24
Layer 9: Proportionality and Transparency	
Documentation	25
IV. SITUATING THE FRAMEWORK: DOCTRINAL AND	
INSTITUTIONAL CONTEXT	26
<i>A. Relationship to Existing Doctrine</i>	26
<i>B. The Sedona Conference and the Development of</i>	
<i>Standards</i>	26
<i>C. The Judicial Engagement Question</i>	27
<i>D. The Deskilling Horizon</i>	29
CONCLUSION.....	31

INTRODUCTION: THE VALIDATION GAP

The doctrinal framework for technology-assisted review (TAR) in eDiscovery rests on a specific epistemic foundation: the existence of binary, verifiable ground truth against which machine outputs can be statistically evaluated. That foundation is not available for the class of interpretive functions that generative artificial intelligence (Gen AI) is now performing in live discovery matters. The question this article addresses is whether the profession has recognized that gap—and, if not, what a validation framework adequate to the epistemic character of Human AI-Assisted Review

(HAR) should require.¹ In 2012, Judge Andrew Peck’s landmark opinion in *Da Silva Moore v. Publicis Groupe* endorsed TAR for eDiscovery, relying on empirical research that Professors Maura Grossman and Gordon Cormack had published the previous year establishing that predictive coding could outperform exhaustive manual review on recall and precision.² That endorsement was the product of a decade of work: the Text REtrieval Conference (TREC) Legal Track’s experimental validation program (pioneered by Jason R. Baron, Maura R. Grossman, and collaborators), the Richmond Journal of Law and Technology study, the development of Grossman and Cormack’s TAR Glossary, and the Sedona Conference’s principled engagement with the technology.³ TAR earned its doctrinal legitimacy because the profession could answer three questions: What does this technology do? How do you know it worked? What should a court require? The answers were empirically grounded and institutionally supported.

More than a decade later, the profession is deploying a fundamentally different class of technology—generative artificial intelligence (Gen AI)—at significant scale in eDiscovery, and it

¹ Neil Cameron, *From TAR to HAR: Are GenAI Discovery Tools Ready for Forensic Scrutiny?*, LEGAL IT INSIDER (Mar. 9, 2026), <https://legaltechnology.com/2026/03/09/from-tar-to-har-report-are-genai-discovery-tools-ready-for-forensic-scrutiny-read-it-here-now/> [<https://perma.cc/4WAR-LJZP>]. The terminology “Human AI-Assisted Review” and the framework developed in this article were first introduced in that practitioner-facing report. *Id.* The present article elaborates the doctrinal and validation arguments more fully and develops the nine-layer framework with the rigor appropriate to a scholarly venue.

² *Da Silva Moore v. Publicis Groupe*, 287 F.R.D. 182, 189-91 (S.D.N.Y. 2012). Judge Andrew Peck’s approval of predictive coding in *Da Silva Moore* is widely regarded as the foundational judicial endorsement of TAR for eDiscovery. See The Sedona Conference, *TAR Case Law Primer, Second Edition*, 24 SEDONA CONF. J. 1, 11 (2023) (describing *Da Silva Moore* as the first published opinion to recognize TAR as an acceptable way to search for relevant ESI).

³ See generally Maura R. Grossman & Gordon V. Cormack, *Technology-Assisted Review in E-Discovery Can Be More Effective and More Efficient Than Exhaustive Manual Review*, RICH. J.L. & TECH., Spring 2011, at 1 [hereinafter Grossman & Cormack, *Richmond Study*] (establishing that technology-assisted review processes can achieve recall and precision exceeding those of exhaustive manual review); Karl Schieneman & Thomas C. Gricks III, *The Implications of Rule 26(g) on the Use of Technology-Assisted Review*, 7 FED. CTS. L. REV. 239 (2013) (setting out the documentation and validation obligations Rule 26(g) imposes on parties using TAR); Maura R. Grossman & Gordon V. Cormack, *The Grossman-Cormack Glossary of Technology-Assisted Review*, 7 FED. CTS. L. REV. 1 (2013) (supplying the field’s standard technical vocabulary for TAR).

cannot answer those questions. Professors Grossman and Cormack, joined by Professor Jason Baron, identified the problem in a paper on the use of large language models (LLMs) in eDiscovery: “LLM tools and protocols have not yet been demonstrated to be as effective as currently recognized methods for legal research, nor for TAR,”⁴ and the validation methodologies that would permit such a demonstration do not yet exist. Yet deployment continues. Multiple major eDiscovery providers have integrated Gen AI into production workflows.⁵ And, as of the time of writing, no sworn declarations have been identified as having been filed in United States courts attesting to the accuracy or validation of Gen AI outputs used in discovery.⁶ The validation architecture that accompanied TAR’s introduction is absent.

This article argues that the absence is not accidental. It reflects a structural mismatch between the validation methodologies developed for TAR and the epistemic character of what Gen AI does in the functions for which it is increasingly deployed. When a Gen AI system generates an early case assessment narrative, produces custodian summaries, constructs a document chronology, identifies thematic clusters, or drafts privilege log entries, it is not performing a classification function that can be evaluated against binary ground truth. It is performing an interpretive cognitive synthesis for which no population-level statistical test currently exists that captures the interpretive adequacy of these outputs in their litigation contexts.⁷ The recall

⁴ Maura R. Grossman, Gordon V. Cormack & Jason R. Baron, *Does the LLMperor Have New Clothes? Some Thoughts on the Use of LLMs in eDiscovery*, NAT’L L. REV. (Nov. 4, 2024), <https://natlawreview.com/node/281097/printable/pdf> [hereinafter Grossman et al., *LLMperor*] (footnote omitted) [<https://perma.cc/N3C2-MA3Z>].

⁵ See, e.g., Video Interview with Brad Kolacinski, Partner, Control Risks (Feb. 19, 2026) (describing Gen AI deployment across multiple investigations); Video Interview with Eric Crawley, Senior Vice President of Legal Sols., Epiq (Feb. 25, 2026) (confirming Gen AI use in large-scale document review programs).

⁶ Video Interview with Jason R. Baron, Professor of the Prac., Coll. of Info. at Univ. of Md. (Feb. 24, 2026). Professor Baron noted that despite widespread Gen AI deployment, he had found no sworn declarations filed in U.S. courts attesting to the accuracy or validation of Gen AI outputs used in discovery—an evidentiary silence with no parallel in the TAR context. *Id.*

⁷ The distinction between classification and interpretation maps onto the information-retrieval distinction between retrieval and summarization tasks. In TAR, the machine performs a classification function evaluable against binary ground truth. In HAR, the machine performs a synthesis function for which no binary ground truth exists, carrying direct implications for the Rule 26(g) certification obligation.

and precision metrics that underpin TAR validation are structurally unavailable for these tasks. This is the validation gap. At bottom, TAR is a retrieval and binary classification task; HAR is primarily to be understood as an interpretive synthesis and summarization task. The two technologies are better conceived as overlapping rather than as wholly distinct: where Gen AI is deployed for binary document categorization—responsiveness, privilege—the statistical validation framework developed for TAR applies in principle, and the existing methodology transfers. It is where Gen AI moves beyond categorization into interpretive synthesis—constructing narratives, summaries, meaning, clustering themes, framing the evidentiary record—that the existing framework has nothing to say, and different ground-truth requirements, different failure modes, and evaluation methodologies that do not transfer come into play.

Figure 1 shows the TAR/HAR relationship as a Venn diagram. Binary categorization tasks occupy the shared zone, where TAR’s statistical validation framework transfers. Interpretive synthesis functions are HAR-only territory, governed by the nine-layer framework.

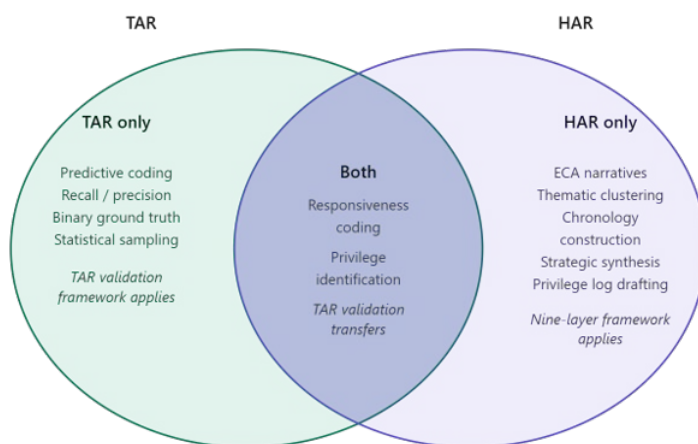


FIGURE 1

What is striking about practitioner accounts of the current deployment environment is not the acknowledgment of governance gaps—that is expected at an early stage of technology adoption—but the absence of any argument that deployment should pause pending validation. Practitioners describe the current state with equanimity, even enthusiasm.⁸ This article does not argue that deployment should cease. It argues that sanguinity about the current state is not a substitute for the institutional work of building a validation architecture, and that the legal profession's interest lies in completing that work before judicial compulsion, not in response to it.

Recent practitioner discussions illustrate the emerging tension. At Legalweek 2026, a mock courtroom exercise examining the use of Gen AI in document review concluded with a ruling by Judge Peck, the author of the original TAR-defining *Da Silva Moore v. Publicis Groupe* judgment, that such workflows could be defensible based primarily on conventional validation metrics such as recall and precision.⁹ That conclusion reflects an understandable

⁸ See sources cited *supra* note 5.

⁹ Legalweek 2026, Mock Courtroom Exercise: Generative AI in Document Review (New York, Mar. 2026) (observed by the author). The presiding judge concluded that Gen

instinct to extend the TAR validation framework to newer tools. But it also illustrates the central difficulty this article addresses: those metrics measure *retrieval accuracy*, not the *interpretive synthesis* through which HAR systems summarize, cluster, and frame the evidentiary record.

The article proceeds as follows. Part I traces the doctrinal development of TAR validation, identifying the epistemic foundations on which the field's statistical methodology rests and the specific claims that methodology can and cannot support. Part II analyzes three positions in the current debate: the continuity position, the reliance-point model, and the discontinuity position. Part III proposes the nine-layer validation architecture. Part IV situates the framework within the broader doctrinal landscape and considers the conditions under which judicial engagement is likely.

I. THE DOCTRINAL ARCHITECTURE OF TAR VALIDATION

A. *The Epistemic Foundations of Recall and Precision*

TAR emerged as a doctrinal category because the legal profession could articulate, with scientific precision, what it was and how to know whether it had worked.¹⁰ Grossman and Cormack's 2011 Richmond Journal of Law and Technology study established a core claim: technology-assisted review processes could achieve recall and precision rates exceeding exhaustive manual review, and those rates could be estimated with statistical confidence from random sampling.¹¹ This claim had a specific epistemic structure. Recall and precision are mathematically defined, empirically testable, and statistically bounded. Given a document population of known size and a sampling protocol of sufficient power, one can make calibrated statements about the probability that a TAR process has achieved adequate recall.

AI-assisted review could be defensible under existing validation frameworks, a conclusion this article examines critically in Part II. *See supra* note 7.

¹⁰ *See generally* Grossman & Cormack, *Richmond Study*, *supra* note 3 (establishing the empirical basis for the statistical validation of TAR).

¹¹ Grossman & Cormack, *Richmond Study*, *supra* note 3, at 28-29, 35-37 (reporting recall and precision results for the two participating teams, whose methods differed—the University of Waterloo applied supervised machine learning, while H5 used a rule-based method—and describing the stratified sampling and statistical inference by which those rates were estimated).

That epistemic structure rests on a fundamental assumption: the availability of an operationally binary reference standard. For a specified request, coding protocol and adjudication process, each document is assigned to the responsive class or the non-responsive class. Relevance in the eDiscovery context is defined by the requests for production and constrained by Rule 26(b) proportionality requirements.¹² Human reviewers can disagree about relevance at the margins, and Grossman and Cormack documented that disagreement extensively.¹³ But the reference standard remains operationally determinate: the protocol and its adjudication process yield a single coding for each document, and the validity of statistical estimation depends on the assumption that the machine's binary output can be compared against human binary judgments.¹⁴

The Schieneman-Gricks framework for applying Rule 26(g) to TAR built directly on this epistemic structure. The framework required parties to document collection, training, stabilization, and validation phases, with statistical testing at each stage.¹⁵ Recall estimation through random sampling became the standard—not a legal formality, but a scientific necessity—because the statistical claims the framework required could only be made if the estimation methodology was sound.¹⁶ Grossman and Cormack's response to the Schieneman and Gricks article refined this framework, pushing back on validation requirements lacking empirical justification while affirming the statistical core.¹⁷ The Special Interest Group on

¹² FED. R. CIV. P. 26(b)(1).

¹³ See Maura R. Grossman & Gordon V. Cormack, *Inconsistent Responsiveness Determination in Document Review: Difference of Opinion or Human Error?*, 32 PACE L. REV. 267, 272-76 (2012).

¹⁴ The "ground truth" problem is central to this article's argument. In TAR, ground truth is established by human review of training documents: a document is relevant, or it is not. In HAR, there is no analogous ground truth for a summary, a timeline, or a thematic characterization. Multiple summaries of the same corpus may be "accurate" yet differ materially in emphasis, framing, and completeness.

¹⁵ Paul W. Grimm, Foreword to Karl Schieneman & Thomas C. Gricks III, *The Implications of Rule 26(g) on the Use of Technology-Assisted Review*, 7 FED. CTS. L. REV. 239, 239-40 (2013).

¹⁶ Video Interview with Maura R. Grossman, Rsch. Professor, Sch. of Comput. Sci. at Univ. of Waterloo (Feb. 25, 2026).

¹⁷ Maura R. Grossman & Gordon V. Cormack, *Comments on "The Implications of Rule 26(g) on the Use of Technology-Assisted Review,"* 7 FED. CTS. L. REV. 285, 287-88 (2014). The debate between this response and Schieneman & Gricks is instructive: the profession benefited from a rigorous empirical argument about what validation

Information Retrieval (SIGIR) 2014 study then established the relative merits of different machine-learning protocols.¹⁸ The profession's engagement with TAR validation was, in this respect, a model of how doctrinal development should interact with empirical research.

B. The Sedona Framework and the Transparency Norm

The Sedona Conference's contribution to TAR's doctrinal legitimacy was normative. The Sedona Principles established that producing parties are best situated to evaluate appropriate review methodologies, and that proportionality—not perfection—is the governing standard.¹⁹ But the Sedona framework also embedded a transparency norm: whatever methodology is used, the producing party must be able to explain it, and the explanation must be capable of meaningful evaluation by courts and opposing counsel.²⁰

This transparency norm is directly implicated by HAR deployment. A summary generated by a Gen AI system can be reproduced and checked against its cited sources. But it cannot be evaluated for *completeness* in the way that a TAR review can be evaluated for *recall*. No sample yields a population-level estimate of interpretive completeness analogous to recall, because there is no reference standard against which a thematic characterization of a large corpus can be scored. Sampling can still provide qualitative and adversarial assurance, as Layer 5 requires; what it cannot provide is a statistical warrant. The transparency norm demands an explanation, but the statistical machinery that made TAR's explanations meaningful is unavailable.

obligations were scientifically justified. No equivalent empirical debate has occurred for HAR.

¹⁸ Gordon V. Cormack & Maura R. Grossman, *Evaluation of Machine-Learning Protocols for Technology-Assisted Review in Electronic Discovery*, PROCS. 37TH INT'L ACM SIGIR CONF. ON RSCH. & DEV. INFO. RETRIEVAL 153, 154 (2014), <https://dl.acm.org/doi/epdf/10.1145/2600428.2609601> [hereinafter Cormack & Grossman, SIGIR 2014] [<https://perma.cc/Q5RH-JVPL>].

¹⁹ *The Sedona Conference, The Sedona Principles, Third Edition: Best Practices, Recommendations & Principles for Addressing Electronic Document Production*, 19 SEDONA CONF. J. 1, 65, 118 (2018) [hereinafter *The Sedona Principles*]. Principle 6 provides that responding parties are best situated to evaluate appropriate methodologies for preserving and producing their own electronically stored information (ESI). *Id.* at 118.

²⁰ *Id.*

The statistical metrics that validated TAR measure the accuracy of document retrieval. They do not measure the completeness or adequacy of the interpretations drawn from those documents.

C. Rule 26(g) and the “Reasonable Inquiry” Standard

Federal Rule of Civil Procedure 26(g)(1) requires attorneys to certify, after “reasonable inquiry,” that discovery disclosures are “complete and correct” and that discovery responses comply with the rules.²¹ The Advisory Committee notes establish that the “reasonable inquiry” standard is objective—measured against what a reasonable attorney in the circumstances would have done—and analogous to the Rule 11 standard.²² Violations can result in mandatory sanctions.²³

Schieneman and Gricks argued that Rule 26(g) imposes specific obligations on parties using TAR: documentation of the review process, transparency about methodology, and demonstration that the process was reasonably designed to achieve complete and accurate production.²⁴ Judge Grimm’s foreword to that article,²⁵ and his earlier opinion in *Mancia v. Mayflower Textile Services Co.*,²⁶ established that the “reasonable inquiry” obligation requires counsel to be actively involved in the discovery process, not merely to delegate and certify.²⁷

The HAR context raises a specific question under Rule 26(g) that has not been addressed in case law or commentary: what does “reasonable inquiry” require when the production strategy is informed by AI-generated interpretive outputs? If an Early Case

²¹ FED. R. CIV. P. 26(g)(1).

²² FED. R. CIV. P. 26(g) advisory committee’s notes to 1983 amendments.

²³ FED. R. CIV. P. 26(g)(3).

²⁴ Schieneman & Gricks III, *supra* note 3, at 248, 273-74.

²⁵ See Grimm, *supra* note 15, at 239-40.

²⁶ *Mancia v. Mayflower Textile Servs. Co.*, 253 F.R.D. 354, 357, 365 (D. Md. 2008) (holding that Rule 26(g) imposes an affirmative obligation of reasonable inquiry on counsel in framing and responding to discovery requests, and that discovery is to be conducted cooperatively rather than adversarially).

²⁷ Video Interview with Paul W. Grimm, Retired U.S. Dist. J., U.S. Dist. Ct. for the Dist. of Md. (Feb. 14, 2026). Judge Grimm described his test for adequate discovery validation, such as asking “Can you demonstrate what you did, why you did it, and what you found?” *Id.* He expressed the view that this test, applied to HAR outputs, requires methodology documentation currently absent from most deployments. *Id.*

Assessment (ECA) narrative generated by a Gen AI system characterizes the dataset in ways that cause counsel to undercollect from certain custodians, the resulting disclosure may not be “complete and correct” even though counsel relied in good faith on an AI output that appeared accurate. The certification obligation is an adequacy-of-process test, and the question of what process is adequate is precisely what the absence of a validation framework for HAR leaves unanswered.²⁸

The central question in the TAR debates was whether the outputs of a technology-assisted workflow could be shown to be statistically reliable. HAR systems shift the focus of that inquiry. Where AI tools perform *interpretive synthesis*—summarizing documents, clustering themes, and shaping the reviewer’s understanding of the evidentiary record—the critical issue is no longer merely the accuracy of the system’s outputs. It is whether the workflow within which those outputs are generated and relied upon can be defended as a reasonable inquiry under the governing rules of discovery. In that sense, the validation problem posed by HAR is not simply an extension of TAR; it reframes the object of validation itself.

II. THREE POSITIONS IN THE CURRENT DEBATE

Scholars and practitioners have begun to advance three distinct responses to this shift in the object of validation. This Part examines each in turn, drawing on original interviews conducted by the author with leading practitioners,²⁹ scholars,³⁰ and jurists.³¹

²⁸ Rule 26(g)(1)(A) provides that an attorney signing a disclosure certifies it is “complete and correct as of the time it is made.” The question this article raises is whether a “reasonable attorney” can certify completeness when the early case assessment on which the collection and production strategy was based relied on AI-generated interpretive outputs that were never validated.

²⁹ Video Interview with Jim Sullivan, CEO, eDiscovery AI (Feb. 26, 2026). Sullivan’s two-rule framework is set out in Part II.B.

³⁰ Video Interview with Maura R. Grossman, *supra* note 16; Video Interview with Jason R. Baron, *supra* note 6. Professor Grossman described current Gen AI deployment as the “Wild West” and stated that the profession has “just skipped step one”—a reference to the pre-deployment validation that TAR received through TREC, the Richmond Study, and the SIGIR research program before *Da Silva Moore*. Video Interview with Maura R. Grossman, *supra* note 16.

³¹ Video Interview with Paul W. Grimm, *supra* note 27.

A. The Continuity Position

The continuity position holds that Gen AI's deployment does not require fundamentally new validation methodologies because the core production obligation remains unchanged: documents are relevant or they are not, and production decisions remain binary. On this view, Gen AI is a more sophisticated tool for performing the same tasks that TAR performed—document classification, privilege identification, relevance ranking—and the existing validation framework, adapted for LLM-based classifiers, is adequate.³²

The continuity position has genuine merits. For binary classification tasks—relevance determination and privilege identification—the statistical validation machinery developed for TAR is applicable in principle.³³ That position is analytically sound within its scope.³⁴ Its limitation is that it describes Gen AI deployment as it should be—confined to binary classification

³² The continuity position is articulated in various forms by major eDiscovery vendors and practitioners. The most rigorous published version is Tara Emory, Jeremy Pickens & Wilzette Louis, *TAR 1 Reference Model: An Established Framework Unifying Traditional and GenAI Approaches to Technology-Assisted Review*, 25 SEDONA CONF. J. 109 (2024). The paper proposes a five-step process framework (Scope, Label Control Set, Iterate Model, Classify, Validate) and argues it applies regardless of whether the underlying engine is discriminative machine learning (ML), expert systems, or Gen AI. *Id.* at 117-18. Critically, the paper expressly limits its scope to first-pass review—binary document tagging for responsiveness—and does not claim to address interpretive functions. *Id.* at 110. See also Gary Foster & David Jackson, *From TAR to GenAI: Rethinking eDiscovery with Large Language Models*, ALVAREZ & MARSAL (June 4, 2025), <https://www.alvarezandmarsal.com/thought-leadership/from-tar-to-genai-rethinking-ediscovery-with-large-language-models> [<https://perma.cc/JG2L-4EEN>] (noting LLMs offer “chain of thought” transparency as an improvement over TAR’s “black box”).

³³ See, e.g., Schieneman & Gricks III, *supra* note 3, at 269-73 (surveying validation methodologies including recall estimation, random sampling, and statistical testing). The core insight is that because TAR outputs are binary, population-level statistical testing is possible. *Id.*

³⁴ A court has now taken that view. In *Schulte v. LinkedIn Corp.*, No. 22-cv-00237-HSG (LB), 2026 WL 1905851 (N.D. Cal. June 30, 2026), ECF No. 203, at 6, Magistrate Judge Beeler recorded that LinkedIn had disclosed its use of Relativity aiR—“a form of Technology Assisted Review”—to filter out non-responsive documents, and held that the disclosure satisfied the TAR-disclosure obligation in paragraph 5(a) of the operative ESI order. The court denied the plaintiffs’ motion to compel further metrics, including elusion estimates, the document error rate and the number of human reviewers validating the tool’s predictions, on the ground that discovery on discovery requires a showing of a specific deficiency rather than speculation. *Id.* at 6-7. The deployment was confined to responsiveness filtering, and for that use the assimilation is correct. The decision is properly read as continuity applied to the case that fits it, not as a judicial resolution of the interpretive question.

tasks—rather than as it is. In practice, Gen AI tools are being deployed as synthesizers of interpretive content: generating ECA narratives, constructing document chronologies, producing thematic summaries. These are not *classification* functions. They are *cognitive synthesis* functions, and the existing validation framework has nothing to say about them. The continuity position’s limitation is therefore one of scope rather than analysis: it correctly describes the validation framework for categorization tasks but does not address the interpretive synthesis functions for which that framework was not designed.³⁵

The most rigorous published articulation of this position is Emory, Pickens, and Louis’s TAR 1 Reference Model, published in the Sedona Conference Journal in 2024. The paper proposes a five-step reference model—Scope, Label Control Set, Iterate Model, Classify, and Validate—and argues that this process framework applies regardless of the underlying algorithmic engine: whether that engine is discriminative supervised machine learning, expert systems, or a Gen AI large language model.³⁶ The engine varies; the process does not. On this view, Gen AI used for first-pass binary classification is not epistemically different from discriminative TAR; it is simply a new engine in the same vehicle, subject to the same process discipline and the same statistical evaluation. The framework of Emory, Pickens, and Louis is analytically careful, and, within its explicitly stated scope, is correct. The paper’s own

³⁵ The limitation is now visible in a stipulated order. The ESI protocol entered in *James v. Cerebras Systems Inc.*, No. 4:25-cv-09361-AMO (N.D. Cal. July 7, 2026), ECF No. 74, defines “AI Responsiveness Review” to include workflows using large-language models, deep-learning classifiers, embedding-based similarity analysis, semantic clustering, predictive redaction systems or generative summarization models, where used to determine whether a document will be produced, withheld or redacted. *Id.* § VII.8. A party electing that review must disclose the identity, version and hosting environment of each system, together with all prompts, templates, instruction sets and parameter configurations used to guide it, with any later prompt modification served in redline within three business days. *Id.* app. 4(a)(i), (c)(v). The validation the protocol requires is nonetheless the apparatus of binary classification: richness, recall, precision and elusion, sampled at a 95% confidence level with a $\pm 2\%$ margin of error, against an expectation that elusion not exceed 3%. *Id.* app. 4(g). The quality-control provisions require the responding party to describe its checks for hallucination and over-summarization, *id.* app. 4(e)(iv)(b), but supply no standard against which either could be assessed. The protocol records the problem without supplying a validation settlement for it. Disclosure of inputs, however comprehensive, is not validation of outputs.

³⁶ Emory et al., *supra* note 32, at 117-18.

limitation is instructive: it is expressly confined to first-pass review—the binary tagging of documents for responsiveness—and explicitly does not address summarization, thematic clustering, chronology construction, privilege log generation, or strategic narrative synthesis. The TAR 1 Reference Model answers the question of how to defensibly use Gen AI for classification. It does not claim to answer the question this article addresses: what happens when Gen AI is used for something other than classification?

B. The Reliance-Point Model

The most sophisticated challenge to the claim that HAR creates a new validation problem comes from what this article terms the “reliance-point model.” Jim Sullivan, CEO of eDiscovery AI, articulates a two-rule framework.³⁷ For Rule 1—binary classification—existing TAR validation methods are “completely appropriate.”³⁸ For Rule 2—everything else—no population-level validation is required because the professional responsibility obligation to verify AI output attaches not at the moment of generation but at the moment of reliance in a formal proceeding.³⁹ If a lawyer independently verifies the source documents before relying on any specific claim in a filing, the accuracy of the intermediate AI output is, in Sullivan’s formulation, “basically irrelevant.”⁴⁰

Sullivan has observed that AI-generated summaries can be comprehensively inaccurate without producing any formal legal consequence beyond, as he put it, “a bad time” in case preparation.⁴¹ The mechanism of the framework is the citation model: every Gen AI output is anchored to a specific source document, and the lawyer’s obligation is to verify that source before relying on the output.

³⁷ Video Interview with Jim Sullivan, *supra* note 29.

³⁸ *Id.*

³⁹ *Id.*

⁴⁰ *Id.*

⁴¹ *Id.* Sullivan stated: “You can create summaries on your entire dataset. They can all be 100% wrong. And there’s no problems—other than you’re going to have a bad time in how you prepare your case.” *Id.*

Within its own assumptions, this framework is coherent. It draws a defensible line between internal case preparation and formal reliance in proceedings. It correctly identifies professional responsibility as the governing framework for the external reliance case. And it provides a practical mechanism for operationalizing that obligation. This article does not dispute the framework within its scope of application. It argues that the framework has three structural vulnerabilities that prevent it from resolving the validation gap.

The first vulnerability is behavioral. The reliance-point model depends on lawyers consistently applying rigorous independent verification before relying on AI-generated summaries in filings. This assumption runs directly against the well-documented phenomenon of automation bias, which is the tendency of human operators to reduce verification rigor in proportion to the fluency and apparent reliability of automated outputs.⁴² The pattern is documented in *Mata v. Avianca*: attorney Schwartz was “operating under the false perception that [ChatGPT] could not possibly be fabricating cases on its own.”⁴³ This is automation bias in its paradigm form. The EU AI Act’s identification of automation bias as a structural governance concern requiring design-level mitigation reflects the same insight: reliance on practitioner vigilance alone, without structured workflow protocols, degrades under real-world conditions.⁴⁴ Automation bias is not a character

⁴² See generally Lisanne Bainbridge, *Ironies of Automation*, 19 AUTOMATICA 775 (1983) (describing the erosion of operator vigilance and skill that accompanies highly reliable automation); Bonnie M. Muir, *Trust in Automation: Part I. Theoretical Issues in the Study of Trust and Human Intervention in Automated Systems*, 37 ERGONOMICS 1905 (1994) (developing a model of human trust in machines, extending models of interpersonal trust to the human-machine relationship, and relating trust to the operator’s choice between manual and automatic control).

⁴³ *Mata v. Avianca, Inc.*, 678 F. Supp. 3d 443, 451, 464-65 (S.D.N.Y. 2023) (sanctioning attorneys for submitting AI-generated fictitious case citations). See also *Johnson v. Dunn*, 792 F. Supp. 3d 1241, 1266-68 (N.D. Ala. 2025) (escalating to a disqualification sanction).

⁴⁴ Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence, art. 14(4)(b), 2024 O.J. (L 1689) 60 (requiring that high-risk AI systems be designed so that the natural persons overseeing them remain alert to the tendency to over-rely on automated output). See generally Johann Laux & Hannah Ruschemeier, *Automation Bias in the AI Act: On the Legal Implications of Attempting to De-Bias Human Oversight of AI*, 16 EUR. J. RISK REGUL. 1519 (2025).

failing; it is a documented, predictable feature of human cognition alongside fluent, high-performing systems, observed in aviation, medicine, and military decision-making long before AI entered legal practice.⁴⁵ The doctrinal implication is direct: Rule 26(g)'s "reasonable inquiry" standard is an objective test, measured against what a reasonable attorney in the circumstances would have done.⁴⁶ A workflow that predictably induces reduced verification vigilance does not become reasonable because individual practitioners intended to check carefully. When a workflow design makes verification failures predictable, the reasonable inquiry obligation may require structural mitigation—embedded protocols, escalation triggers, logged overrides—rather than reliance on individual caution alone.⁴⁷

The second vulnerability concerns the scope of the citation model itself. The model works well for AI outputs structurally anchored to specific source documents. It does not generalize to interpretive outputs at higher levels of abstraction: thematic characterizations of a large corpus, strategic narrative synthesis across thousands of documents, or issue identification that emerges from pattern recognition rather than specific document content. For these outputs, there is no source document to verify. The citation model's verification mechanism is unavailable, and the reliance-point trigger either never fires or fires at a point where the distortion has already cascaded through the production strategy.

There is a subtler limitation that applies even to well-cited, document-anchored outputs—one that Grossman has identified as an unresolved precision-recall problem at the interpretive level.⁴⁸ A citation confirms that a document exists and was consulted; it does not confirm that the summary captured all legally relevant content within it. A summary may correctly cite a document while silently omitting a passage that changes its significance. The citation directs counsel to the correct document but provides no indication

⁴⁵ See generally Bainbridge, *supra* note 42 (documenting the phenomenon in aviation and industrial process control).

⁴⁶ See *supra* notes 21-22 and accompanying text.

⁴⁷ See Herbert B. Dixon Jr. et al., *Navigating AI in the Judiciary: New Guidelines for Judges and Their Chambers*, 26 SEDONA CONF. J. 1, 4 (2025). The Guidelines identify automation bias directly: "Users must exercise vigilance to avoid becoming 'anchored' to the AI's response." *Id.*

⁴⁸ Video Interview with Maura R. Grossman, *supra* note 16.

of what the summary may have omitted, and the AI's interpretive framing of that document has, by that point, already structured counsel's expectations about what it contains. This is a precision-recall problem at the interpretive level rather than the classification level—not whether the *right documents* were retrieved but whether the *right meaning* was extracted from them. Sullivan's citation model does not resolve this because verification of existence is not the same as verification of completeness.⁴⁹ If the burden of catching interpretive omissions requires re-reading the entire source document, the efficiency rationale for AI-generated summaries collapses, and the cognitive distortion introduced by the AI's framing may make that re-reading less rigorous than it would have been without the summary. This distinction—interpretive omission—is structurally different from the retrieval failures that TAR validation is designed to detect. A retrieval failure leaves a document uncollected; an interpretive omission leaves the document collected but leaves its legally significant content unregistered, framed away, or subordinated to the AI's preferred characterization. The cognitive distortion introduced by that framing may persist even when counsel subsequently reads the underlying document because the AI's account has already anchored expectations about what the document contains and what significance it bears.

If verification of completeness requires counsel to re-read the entire underlying document independently, the efficiency rationale for AI-generated summaries collapses. The verification step necessary to ensure interpretive adequacy is functionally equivalent to the review the summary was intended to replace.

The third vulnerability is the internal/external boundary problem. Sullivan's framework draws a clean analytical line between internal case preparation and external formal reliance.⁵⁰ But this line is operationally porous. AI-generated interpretive framing shapes which documents are pulled for review, which custodians are prioritized, which arguments are developed, and what is produced or withheld. If that framing is systematically distorted, the consequences may never surface in a formal

⁴⁹ *Id.*

⁵⁰ Video Interview with Jim Sullivan, *supra* note 29.

proceeding.⁵¹ The reliance-point trigger never fires, but the distorted framing has already shaped the production, the witness strategy, and the settlement calculus. The harm is real and legally material; it is simply not the harm that the formal reliance framework addresses. The boundary between work product and discoverable AI-generated materials adds a further dimension: where AI interpretive outputs are embedded in counsel's case preparation, whether those outputs are shielded by work-product protection or subject to discovery when their accuracy is placed in issue remains doctrinally unresolved. The answer will shape how the internal/external boundary problem is ultimately litigated.

C. The Discontinuity Position and the Medicine Analogy

The discontinuity position, articulated most forcefully by Professor Grossman, holds that Gen AI's deployment in eDiscovery without adequate validation is the professional equivalent of deploying a pharmaceutical without clinical trials.⁵² The TREC Legal Track was proposed by Professor Baron together with Doug Oard and David Lewis, and drew on Baron's experience searching White House email at the National Archives and Records Administration (NARA). It served as the eDiscovery equivalent of clinical trials for TAR: a systematic, independent, empirical validation program that established efficacy before judicial endorsement.⁵³ No equivalent program exists for HAR.

⁵¹ FED. R. CIV. P. 26(g)(1)(A) governs disclosures—which are external—while FED. R. CIV. P. 26(g)(1)(B) governs discovery requests, responses, and objections. Internal case preparation is not, on its face, governed by either provision. For the boundary between work product and discoverable materials in the AI context, see Tremblay v. OpenAI, Inc., No. 23-cv-03223, 2024 WL 3748003, at *2-3 (N.D. Cal. Aug. 8, 2024) and Scott A. Milner & Matthew J. Hamilton, *When AI Meets Privilege: Early Court Decisions*, MORGAN LEWIS (Feb. 18, 2026), <https://www.morganlewis.com/pubs/2026/02/when-ai-meets-privilege-early-court-decisions> [<https://perma.cc/3BZJ-JSUH>].

⁵² Video Interview with Maura R. Grossman, *supra* note 16 (analogizing current Gen AI deployment to pharmaceutical approval: the FDA tests before approving a drug, whereas the eDiscovery profession has “just skipped step one”). The medicine analogy is elaborated through discussion of the TREC Legal Track as the eDiscovery equivalent of clinical trials. *Id.*

⁵³ Video Interview with Jason R. Baron, *supra* note 6 (describing the origins of the TREC Legal Track and its relationship to Baron's work searching White House email at NARA); see also Grossman & Cormack, *Richmond Study*, *supra* note 3, at 25-29.

Grossman, Cormack, and Baron’s 2024 paper frames the question as one of epistemology, not merely practice.⁵⁴ The tasks involved in eDiscovery require case-specific knowledge—specific names, dates, filings, legal issues—that LLMs, trained on general corpora, are not optimally designed to supply.⁵⁵ The paper asks whether LLMs can be substituted for current search methods and concludes that this question has not been adequately tested.⁵⁶ Identifying a validation gap is necessary; specifying what would close it is the work that remains.

Grossman also identifies a significant concern that the continuity argument tends to underweight: function creep.⁵⁷ A Gen AI system validated for one discovery task is not thereby validated for a different task presenting distinct interpretive challenges and different failure modes. “Just because you tested it for purpose A doesn’t mean it’s valid for purpose B,” Grossman has observed⁵⁸—validity for relevance classification does not imply validity for privilege analysis, and validity for privilege analysis does not imply validity for chronology construction or issue tagging. This concern is not hypothetical. Vendors routinely deploy the same Gen AI infrastructure across multiple discovery functions, each of which presents different epistemic demands. The absence of task-specific validation protocols means that the assurance provided by any single validated deployment does not travel across use cases.

III. THE OBJECT-OF-PROOF PROBLEM AND THE NINE-LAYER FRAMEWORK

A. *The Causation Proposition*

The central doctrinal claim this article advances is that HAR changes the *object of proof* in the discovery defensibility inquiry. Under the TAR framework, the object of proof is the adequacy of

⁵⁴ Grossman et al., *LLMperor*, *supra* note 4.

⁵⁵ Grossman et al., *LLMperor*, *supra* note 4 (posing the central question: “Are LLMs able to answer [eDiscovery classification questions] as well as or better than existing practices . . . ?”). The paper’s conclusion is that “LLM tools and protocols have not yet been demonstrated to be as effective as currently recognized methods for legal research, nor for TAR.” *Id.* (footnote omitted).

⁵⁶ *Id.*

⁵⁷ Video Interview with Maura R. Grossman, *supra* note 16.

⁵⁸ Video Interview with Maura R. Grossman, *supra* note 16.

the classification process: the party that has used TAR must demonstrate, through statistical evidence, that its review achieved adequate recall and precision on the population of responsive documents. The inquiry is backward-looking, population-level, and statistically testable.⁵⁹

Under the HAR framework, the object of proof shifts. For interpretive outputs—the ECA narrative that frames the case, the thematic summary that characterizes the dataset, the strategic identification of key custodians—the relevant question is not recall and precision on a population of documents but the *interpretive accuracy* of a synthesis that may shape every subsequent decision in the matter. That question cannot be answered by statistical sampling. It requires a different methodology: structured adversarial review, documented senior-lawyer sign-off, source-anchored verification where possible, and transparency about what methodology was used and why it is adequate.

This shift in the object of proof has direct implications for Rule 26(g) certification. A lawyer certifying that a disclosure is “complete and correct” is, in the HAR context, certifying not just that the classification was adequate but that the interpretive framing shaping the production strategy was adequately verified. The existing case law offers no guidance on what that certification requires. Judge Grimm’s own test (the “Grimm test”)—“Can you demonstrate what you did, why you did it, and what you found?”—applied to a HAR-assisted production workflow, requires methodology documentation that current practice does not supply.⁶⁰

Discovery doctrine has long evaluated review methodologies not by the intrinsic properties of the tools employed but by the reasonableness of the process through which counsel conducts inquiry and certifies discovery responses. The validation architecture proposed here follows the same logic. Rather than attempting to certify the correctness of generative models in the abstract, it evaluates the defensibility of the workflow within which those models are deployed.

⁵⁹ See Schieneman & Gricks III, *supra* note 3, at 269-73.

⁶⁰ Video Interview with Paul W. Grimm, *supra* note 27.

B. A Nine-Layer Validation Architecture

The following framework is offered as a normative proposal for what adequate validation of HAR-assisted productions should require. It is a prescription for developing practice, intended to inform professional standards, judicial guidance, and future empirical research.⁶¹ The nine layers address the full lifecycle of Gen AI deployment in eDiscovery, from initial function mapping through final transparency documentation. They are designed to be proportionate: the level of validation appropriate to complex multi-jurisdictional litigation differs from that appropriate to simpler commercial disputes. Layers 1 and 2 apply to every output and perform the categorization on which the rest depends; Layer 3 addresses binary outputs by extending the existing TAR framework without significant modification; Layers 4–9 address the interpretive synthesis functions for which no existing framework is adequate.

This focus on workflow defensibility reflects the structure of discovery doctrine itself. The obligation imposed by Rule 26(g) is not to guarantee the correctness of every discovery response. It is to certify that counsel has conducted a **reasonable inquiry** into the factual basis of those responses.

The validation problem posed by HAR, therefore, becomes a question of process: whether the AI-assisted workflow through which documents are analyzed and interpreted can be defended as a reasonable inquiry within the meaning of Rule 26(g).

Layer 1: Function Mapping and Use Case Categorization

Before deploying any Gen AI tool, counsel should document the specific functions the tool will perform and categorize each output as either binary (classification, privilege determination, responsiveness coding) or interpretive (summary, narrative, thematic analysis, chronology, privilege log narrative). This categorization determines which validation methodology applies. For binary outputs, the existing TAR validation framework applies

⁶¹ The nine-layer validation architecture proposed in this article is a normative framework, not a description of current practice. Each layer is developed in Part III. The layers are designed to be proportionate to matter complexity and the degree to which AI-generated outputs shaped the production strategy.

without modification: the statistical methodology developed by Grossman, Cormack, and the Sedona Conference, and operationalized most rigorously in Emory, Pickens, and Louis' TAR 1 Reference Model⁶²—and Layer 3 specifies what that requires. For interpretive outputs, the TAR framework does not reach, and Layers 4–9 specify what defensible validation requires in its place. The framework's novel contribution lies in those later layers; Layers 1–3 are an explicit adoption of what the profession already accepts.

Layer 2: Output Type Characterization

For each AI-generated output, counsel should characterize its epistemic type: does it produce a classification, a summary, a narrative, a thematic cluster, or a strategic recommendation? Each type carries different verification obligations. Classifications can be statistically tested. Summaries and narratives must be verified against source documents. Thematic clusters and strategic recommendations require adversarial senior-lawyer review.

Layer 3: Statistical Validation for Binary Outputs

For binary outputs—relevance classification, privilege identification, responsiveness coding—the statistical validation framework developed for TAR applies without modification.⁶³ This layer is the province of Grossman and Cormack's methodology: recall estimation through random sampling, precision measurement, and documentation sufficient to satisfy the Schieneman and Gricks' framework. The continuity position is correct that this layer does not require fundamental revision.

The Emory, Pickens, and Louis' TAR 1 Reference Model provides the most rigorous published specification of what this layer requires: the five-step process framework of Scope, Label Control Set, Iterate Model, Classify, and Validate, applied consistently regardless of the underlying algorithmic engine.⁶⁴ For

⁶² Emory et al., *supra* note 32, at 117-18.

⁶³ See generally Schieneman & Gricks III, *supra* note 3 (specifying the documentation and statistical validation obligations Rule 26(g) imposes on a TAR-assisted production); Cormack & Grossman, *SIGIR 2014*, *supra* note 18 (establishing the relative effectiveness of competing machine-learning protocols).

⁶⁴ Emory et al., *supra* note 32, at 110, 117-18.

binary classification tasks, that framework is adequate, and this article endorses it for Layer 3 deployments.

Layer 4: Source Document Integrity Verification

For AI-generated outputs anchored to specific source documents—privilege log entries, timeline events, custodian summaries—counsel should implement Sullivan’s citation model: every output is linked to its source document, and a designated reviewer verifies the source before any specific claim is relied upon in a filing.⁶⁵ The key implementation requirement is workflow design: the citation model must be structurally embedded in the review process, not left to individual practitioner vigilance.

Layer 5: Contextual Coherence Sampling

For all AI-generated outputs, counsel should implement a sampling protocol designed to identify systematic errors. For interpretive outputs, contextual coherence sampling requires a senior reviewer to independently sample a stratified subset of the documents from which an AI-generated summary was derived, compare the AI output against that sample, and document the findings. Sample size should be proportionate to the complexity of the matter and the centrality of the AI output to the production strategy. This layer applies to all interpretive outputs, including document-anchored ones; Layer 7 addresses the harder case of outputs with no source-document anchor at all, where the citation model’s verification mechanism is wholly unavailable, and Layer 5 sampling alone cannot substitute for the structured adversarial review that Layer 7 requires.

Layer 6: Pre-Deployment Independent Validation

For AI tools being deployed in significant litigation, Grossman’s “step one” should be taken before deployment on the live dataset: independent validation of the tool’s performance on a

⁶⁵ Video Interview with Jim Sullivan, *supra* note 29. Sullivan described the citation model as the mechanism by which his platform operationalizes the professional responsibility obligation: every AI-generated output is anchored to a source document, and the lawyer verifies that source before relying on any specific claim in a filing.

test corpus approximating the matter's document population.⁶⁶ This is the pharmaceutical equivalent of Phase II clinical trials—not full empirical validation but a structured pre-deployment test establishing a baseline of expected performance. Counsel should document the test methodology, corpus, evaluation criteria, and findings. This documentation forms part of the transparency record required by Layer 9. Layer 6 is the framework's direct response to the discontinuity position's core critique: the profession has deployed Gen AI without the pre-deployment empirical testing that TAR received through the TREC Legal Track, and the Richmond Journal study before judicial endorsement in *Da Silva Moore*.

Layer 7: Source-Anchored Interpretive Verification

For interpretive outputs not anchored to specific source documents—thematic cluster characterizations, strategic narrative syntheses, corpus-level issue identifications—the citation model is unavailable, and a different verification methodology is required. “Source-anchored” here describes the verification method rather than the output: an independently selected corpus sample supplies the evidential anchor that the citation model cannot. This layer requires a structured adversarial review by a senior lawyer who has independently sampled the underlying corpus, comparing the AI-generated characterization against that independent sample, documenting any divergences, and providing a signed statement of findings that is specific enough to be meaningful.

The key design principle is that this layer must transform “a lawyer reviewed it” from a comfort phrase into an accountable method. The review must be documented; the methodology must be described; and the sign-off must be tied to specific findings rather than general satisfaction. This is what makes the verification meaningful for Rule 26(g) certification purposes.

Layer 8: Structured Human Override Protocols

This layer exists to make the verification obligations of the preceding layers operationally reliable in the face of automation bias. Structured human override protocols should include: mandatory citation checking before any AI-generated claim is

⁶⁶ Video Interview with Maura R. Grossman, *supra* note 16.

incorporated in a filing, escalation triggers requiring senior review when AI-generated outputs diverge from counsel's independent assessment, logged override records documenting when and why an AI-generated output was rejected or modified, and periodic adversarial sampling reviews at fixed intervals in extended reviews.⁶⁷ These protocols are not alternatives to professional responsibility; they are mechanisms that make professional responsibility operationally reliable rather than aspirationally stated. The motivating principle is what may be termed the paradox of reliability, following Bainbridge's account of the ironies of automation: the more consistently a system performs in ordinary cases, the stronger the cognitive pull toward accepting its outputs in the exceptional case where it is wrong.⁶⁸ A lawyer who has verified one hundred AI-generated summaries and found them accurate is not better positioned to catch the hundred-and-first error; she is worse positioned because the prior track record actively suppresses skepticism. Structural protocols exist precisely to counteract that dynamic.⁶⁹

Layer 9: Proportionality and Transparency Documentation

Counsel should maintain documentation of the entire AI deployment process sufficient to satisfy the Grimm test: Can you demonstrate what you did, why you did it, and what you found?⁷⁰ This documentation should include: the function mapping from Layer 1, the output type characterization from Layer 2, validation methodology and results from Layers 3 through 8, deviations from the standard workflow and their justifications, and a senior lawyer certification that the documentation accurately reflects the process. The extent of that record is not fixed. It scales with the complexity of the matter, the sensitivity of the AI output, the centrality of that output to the production strategy, and the consequence of error, so that a single checked interpretive output in a small deployment attracts a lighter record than a corpus-level synthesis on which a

⁶⁷ See Bainbridge, *supra* note 42, at 775-76 (describing the paradox that the more reliably an automated system performs, the less equipped its human supervisor becomes to detect the occasions on which it fails).

⁶⁸ *Id.*

⁶⁹ See Dixon et al., *supra* note 47, at 4.

⁷⁰ Video Interview with Paul W. Grimm, *supra* note 27.

served response depends. Courts should not require a full Daubert-style hearing in every case, but counsel should be in a position to produce the documentation if the adequacy of the process is challenged.⁷¹

IV. SITUATING THE FRAMEWORK: DOCTRINAL AND INSTITUTIONAL CONTEXT

A. Relationship to Existing Doctrine

The nine-layer framework does not require new rules. It is an application of existing doctrine—the Rule 26(g) “reasonable inquiry” standard, the Sedona transparency norm, and the professional responsibility obligations of competence and supervision—to a context in which those obligations have not yet been specified.⁷² The doctrinal contribution this article advances is the identification of the *object-of-proof shift* as the mechanism by which HAR creates new validation obligations. A party challenging a HAR-assisted production can argue under existing doctrine that the producing party’s Rule 26(g) certification was inadequate because counsel could not demonstrate that the interpretive framing of the AI-generated ECA narrative was adequately verified. The framework proposed here specifies what adequate verification looks like.

B. The Sedona Conference and the Development of Standards

The most important institutional development in this space is the Sedona Conference’s formation of Working Group 13 on Artificial Intelligence and the Law, launched in January 2025 and charged with coordinating AI-related work across existing Working Groups and developing principles and best practices for AI in legal contexts.⁷³ The nine-layer framework proposed here is offered as a

⁷¹ See Schieneman & Gricks III, *supra* note 3, at 269-73.

⁷² FED. R. CIV. P. 26(g)(1). See generally The Sedona Conference, *supra* note 19; A.B.A. Comm. on Ethics & Pro. Resp., Formal Op. 512 (2024) (providing the first comprehensive ABA guidance on lawyers’ use of generative AI and grounding obligations in Model Rule 1.1 (competence) and 5.3 (supervisory responsibility)).

⁷³ The Sedona Conference’s Working Group 13 on Artificial Intelligence and the Law launched in January 2025, charged in part with coordinating AI-related work across existing Working Groups. See The Sedona Conference, *Working Group 13 on Artificial Intelligence and the Law*, <https://thesedonaconference.org/wgs/wg13>

contribution to the normative development that Working Group 13 is undertaking, not as a substitute for it. The Sedona process has historically translated technical and empirical research into principled guidance that courts and practitioners can apply; the framework here provides the normative architecture on which that guidance can be built.

C. The Judicial Engagement Question

The conditions under which courts are likely to engage with HAR validation are worth examining. TAR received judicial attention because discovery disputes arose in which parties challenged each other's methodologies, and judges—including Judge Grimm and Judge Peck—had both the technical sophistication and the institutional opportunity to develop doctrine.⁷⁴ HAR is likely to attract judicial attention through two mechanisms.

The first is the existing AI-sanctions pathway. *Mata v. Avianca, Inc.* and its successors have established that courts will sanction counsel for submitting AI-generated content without adequate verification.⁷⁵ The reach of that pathway is no longer confined to pleadings. In *United States v. Heppner*, Judge Rakoff held—on what he characterized as a question of first impression nationwide—that documents a criminal defendant generated using a consumer Gen AI platform were protected by neither attorney-client privilege nor the work-product doctrine.⁷⁶ Although these

[<https://perma.cc/5P37-JZEU>] (last visited Aug. 21, 2026) (setting out the Working Group's mission and recording its inaugural meeting on January 16-17, 2025). Working Group 1 has since formed a brainstorming group on the discoverability and admissibility of Gen AI-related data—grounding and training data, prompts, inputs, and outputs—including the data generated when Gen AI tools are used in document review to classify documents for responsiveness and privilege. See The Sedona Conference, *WG1 Call for Volunteers: Brainstorming Group on Discoverability, Admissibility of Gen AI-Related Data*, https://www.thesedonaconference.org/WG1_CFV-GENAI [<https://perma.cc/9QPE-JC4C>] (last visited Aug. 21, 2026).

⁷⁴ See *Da Silva Moore v. Publicis Groupe*, 287 F.R.D. 182 (S.D.N.Y. 2012) (approving the use of predictive coding in eDiscovery); *Mancia v. Mayflower Textile Servs. Co.*, 253 F.R.D. 354 (D. Md. 2008) (setting out counsel's Rule 26(g) obligations of reasonable inquiry and cooperation).

⁷⁵ *Mata v. Avianca, Inc.*, 678 F. Supp. 3d 443, 448-49, 464 (S.D.N.Y. 2023).

⁷⁶ *United States v. Heppner*, 820 F. Supp. 3d 292, 296-99 (S.D.N.Y. 2026). Judge Rakoff ruled, on what he characterized as a question of first impression nationwide, that materials a criminal defendant generated using a consumer version of Anthropic's

cases involve AI use in pleadings and in privileged client communications respectively, the doctrinal logic extends to the discovery setting this article addresses: the Rule 26(g) certification obligations require adequate verification of AI-generated content before it is incorporated in a production. *Johnson v. Dunn*'s escalation from monetary sanctions to disqualification signals judicial impatience.⁷⁷

The second mechanism is structural. As HAR tools become embedded in large-scale litigation, a high-profile case will eventually surface in which a significant production failure is traceable to an error in an AI-generated ECA narrative or thematic characterization. That case will create the opportunity for the doctrinal development this article's framework anticipates. The profession's interest is in building the validation architecture before that case, not in reaction to it. As Grossman observed in an interview, "something is going to go wrong; the question is whether the profession has a framework in place before it does."⁷⁸ The nine-layer architecture proposed in this article is an attempt to ensure that when that case arrives, the answer is yes.

A further indication that the disclosure gap is not merely prospective comes from practitioners already deploying these systems. Senior practitioners at leading eDiscovery providers report that the concern most consistently raised by sophisticated clients is not data leakage, model failure, or cost overrun; it is explainability.⁷⁹ Clients who struggled to internalize recall and precision as metrics for TAR now face a technology they find even

Claude AI were not protected by attorney-client privilege or the work-product doctrine. *Id.* The court rested its privilege ruling on three independent grounds: (i) Claude is not an attorney, so the communications were not between a client and counsel; (ii) Anthropic's privacy policy reserved rights to disclose user data to third parties, defeating any reasonable expectation of confidentiality; and (iii) the documents were not prepared at the direction of counsel. *Id.* at 296-97. The court left open, in dictum, the possibility that AI use directed by counsel under a Kovel-type arrangement might qualify for protection. *Id.* at 297. The HAR validation context is distinct from the consumer-tool facts of *Heppner*, but the doctrinal logic transfers: where Gen AI-generated materials inform counsel's production strategy, the workflow design must be capable of demonstrating the conditions under which counsel directed and supervised that use—a documentation requirement the nine-layer framework is designed to satisfy.

⁷⁷ *Johnson v. Dunn*, 792 F. Supp. 3d 1241, 1266-67 (N.D. Ala. 2025).

⁷⁸ Video Interview with Maura R. Grossman, *supra* note 16.

⁷⁹ Video Interview with Raj Chandrasekar, Chief Tech. & Innovation Officer, Consilio (Feb. 27, 2026).

harder to interrogate. If the practitioners deploying these systems cannot explain them to the lawyers who must certify their outputs under Rule 26(g), the gap between operational deployment and documentable reasonable inquiry is not a future problem. It is already present.⁸⁰

D. The Deskilling Horizon

The validation concerns this article addresses are immediate: they arise the first time a production strategy is shaped by an AI-generated ECA narrative that counsel cannot defend. But, practitioners engaged with HAR workflows have identified a longer-horizon risk that the nine-layer framework addresses only partially and which the profession will need to confront directly: the progressive deskilling of legal judgment through systematic overreliance on AI-generated interpretive outputs.⁸¹

The concern is not that individual lawyers will fail to verify AI outputs in particular matters—that is the automation bias problem already addressed in Layers 7 and 8. The concern is institutional and cumulative. Lawyers have historically developed case knowledge through first-pass document review: the act of reading large volumes of documents generated the contextual understanding of the evidentiary record on which strategic judgment depended. As Kolacinski observed, “you used to get smart on a case by doing doc review.”⁸² If AI-generated summaries and thematic clusters systematically replace that contact with the underlying material, the interpretive capacity that made verification meaningful may itself atrophy—not in any single matter but across a generation of practitioners who have never developed it. A profession that has lost the habit of reading source

⁸⁰ Interviews conducted by the author with senior practitioners at Consilio, Epig, OpenText, Reveal, Control Risks, Everlaw, and QuisLex (February 2026). Several of these practitioners spoke on a non-attribution basis; those who consented to attribution are identified individually elsewhere in these notes. Across these interviews, the concern most consistently identified as paramount among sophisticated clients was not data security or cost but the inability of deploying practitioners to explain, in terms a certifying lawyer could apply, how AI-generated outputs had been validated. This observation informed the article’s analysis of the explainability gap in Part IV.C.

⁸¹ See generally Bainbridge, *supra* note 42 (documenting skill attrition in automated environments).

⁸² Video Interview with Brad Kolacinski, *supra* note 5.

documents is not well-positioned to catch the errors in AI-generated summaries of those documents, however rigorous its verification protocols are on paper.

This concern has a direct doctrinal dimension. Rule 26(g)'s "reasonable inquiry" obligation is an objective standard, measured against what a reasonable attorney in the circumstances would have done. What counts as reasonable inquiry is not fixed: it is calibrated against the profession's collective understanding of adequate practice. If systematic deskilling shifts the baseline of what attorneys actually do, the standard risks being calibrated downward in ways that do not serve the parties whose interests it is designed to protect. The nine-layer framework addresses this partially: Layers 7 and 8 are designed precisely to force lawyers back into structured contact with underlying material, making verification an accountable method rather than a comfort phrase. But the framework addresses the workflow problem in individual matters; it does not address the multi-year institutional question of how the profession maintains the interpretive capacity on which meaningful oversight of AI-generated outputs depends. That question deserves sustained attention in the professional responsibility literature and in the curriculum design decisions of law schools now training the lawyers who will certify AI-assisted productions for the next three decades. Both the deskilling risk and the urgent question of how attorneys should be trained to use AI tools effectively—not merely fluently but critically—are live and pressing concerns. They are, however, outside the direct scope of this article, which addresses the validation of AI-assisted workflows in discovery rather than the formation of the professional competence on which that validation ultimately depends. They are noted here because they constitute the longer horizon within which the nine-layer framework will either succeed or prove insufficient.⁸³

⁸³ The deskilling concern maps onto a broader literature on automation and professional competence. See generally Bainbridge, *supra* note 42 (documenting skill attrition in automated environments as a general phenomenon). For an early articulation of the deskilling concern specific to legal AI, see Video Interview with David Stanton, Partner, Pillsbury Winthrop Shaw Pittman LLP (May 27, 2026). Stanton identified the structural character of the risk with precision: "We have a process that's opaque, and we only see the output . . . We've trusted a process of human beings doing stuff—you sort of trust that the associate has done a good job, but here, we have a process

CONCLUSION

Gen AI tools are being deployed in eDiscovery at significant scale without a validation architecture adequate to their epistemic character. The profession possesses the institutional knowledge to build that architecture; TAR's trajectory from empirical research to judicial endorsement to professional standards demonstrates how technology validation should proceed. The TAR 1 Reference Model recently published in the Sedona Conference Journal demonstrates that the profession can build rigorous process frameworks for AI-assisted classification; the task that remains is building the equivalent for *interpretive synthesis*.⁸⁴ The structural problem is not institutional incapacity but categorical misapplication. The profession is applying a validation framework designed for binary classification to technology performing interpretive cognitive synthesis without examining whether the framework is fit for that purpose.

This article has argued that it does not. HAR performs interpretive cognitive synthesis, not binary document classification, and the statistical validation framework developed for TAR does not, without significant modification, supply a sufficient method for evaluating interpretive outputs. The reliance-point model offers a sophisticated response to this gap—validation obligations attach at the point of formal reliance, not at the point of generation—but has three structural vulnerabilities: automation bias, the citation model's limits for high-abstraction interpretive outputs, and the internal/external boundary's operational porosity.

that's opaque." *Id.* On completeness and validation, he observed: "That's not enterprise-grade validation." *Id.* On the steering-wheel question—whether lawyers will cede interpretive judgment to AI systems entirely—his formulation was: "I'm not ready to take my hands off the steering wheel." *Id.* The training dimension of the deskilling concern is addressed in Neil Cameron, *Legora Launches Legal AI Scholars Program with Nine US Law Schools*, LEGAL IT INSIDER (Mar. 25, 2026), <https://www.legaltechnology.com/latest-news/legora-launches-legal-ai-scholars-program-with-nine-us-law-schools/> [<https://perma.cc/ZH53-UVJE>] (noting that law graduates "fluent in [professional AI tools] but unable to interrogate [their] outputs critically will be more dangerous, not less, than those who never used [them]" and that AI training, while necessary, does not fill the pre-existing gap in practical legal competencies). A.B.A. Comm. on Ethics & Pro. Resp., Formal Op. 512 (2024) addresses competence obligations under Model Rule 1.1 but does not engage with the cumulative institutional dimension of AI-induced skill attrition.

⁸⁴ Emory et al., *supra* note 32.

The nine-layer framework proposed here is intended to close those vulnerabilities by specifying what adequate validation of HAR-assisted productions requires under existing doctrine. More specifically, it preserves TAR's validation discipline while supplying different methods for the interpretive synthesis functions TAR was not designed to perform; it operationalizes the Rule 26(g) "reasonable inquiry" standard in the HAR context by specifying what process documentation that obligation requires, and it provides the Sedona Conference's Working Group 13 with a normative framework on which principled guidance for courts and practitioners can be constructed.

Grossman, Cormack, and Baron's 2024 paper asked whether Gen AI tools have been adequately validated for the functions to which the profession is deploying them.⁸⁵ The answer this article offers is that they have not—not because those tools lack genuine utility but because the validation architecture their interpretive functions require has not been built. Gen AI performs functions in eDiscovery that TAR could not perform, and the profession's task is not to resist that deployment but to construct the methodological infrastructure on which its defensibility depends. The nine-layer framework proposed here is a specification of what that infrastructure should contain. The normative work remains to be done; this article has attempted to define its structure.

⁸⁵ Grossman et al., *LLMperor*, *supra* note 4.