

Narrative Is Not Custody: Why AI Compliance Needs Traversable Rules

A fluent explanation is not a record. A regulated decision needs the rule that applied, the facts that matched it, the exception that was or was not taken, and the named human who answered for the call.

Compliance officers do not buy fluency. They buy a trail. A regulator asking why a transfer was allowed, why a claim was denied, or why a recommendation overrode a guideline is not asking for a probability. The question concerns the rule that applied, the facts that matched it, the exception path that was or was not taken, and the human who stood behind the call. Pure generative systems are poorly suited to that job, even when they sound excellent doing it.

A method that pairs every fact with a tactic and every tactic with a measurable outcome, which the author calls [Factics](#), sets the standard for this problem (Puglisi, 2012). Every fact must lead to a tactic, and every tactic must leave evidence. AI output that cannot meet that standard has not yet become evidence. What turns it into evidence is structure around the parts of a decision that must stay checkable, and neuro-symbolic and knowledge-graph architectures are one route to that structure. Learning handles messy language and pattern recognition, while symbols, graphs, and solvers can hold the obligations, the exceptions, and the proof.

What does a regulated decision owe the people who audit it?

The accountability question comes first and the architecture comes second. As the author defines them, Ethical AI asks whether something should be done. Responsible AI asks who answers when it fails, and it answers with internal controls in which the machine checks the machine. AI Governance asks who decides, by what authority, and at what checkpoint, and it answers with a named human whose authority binds the outcome.

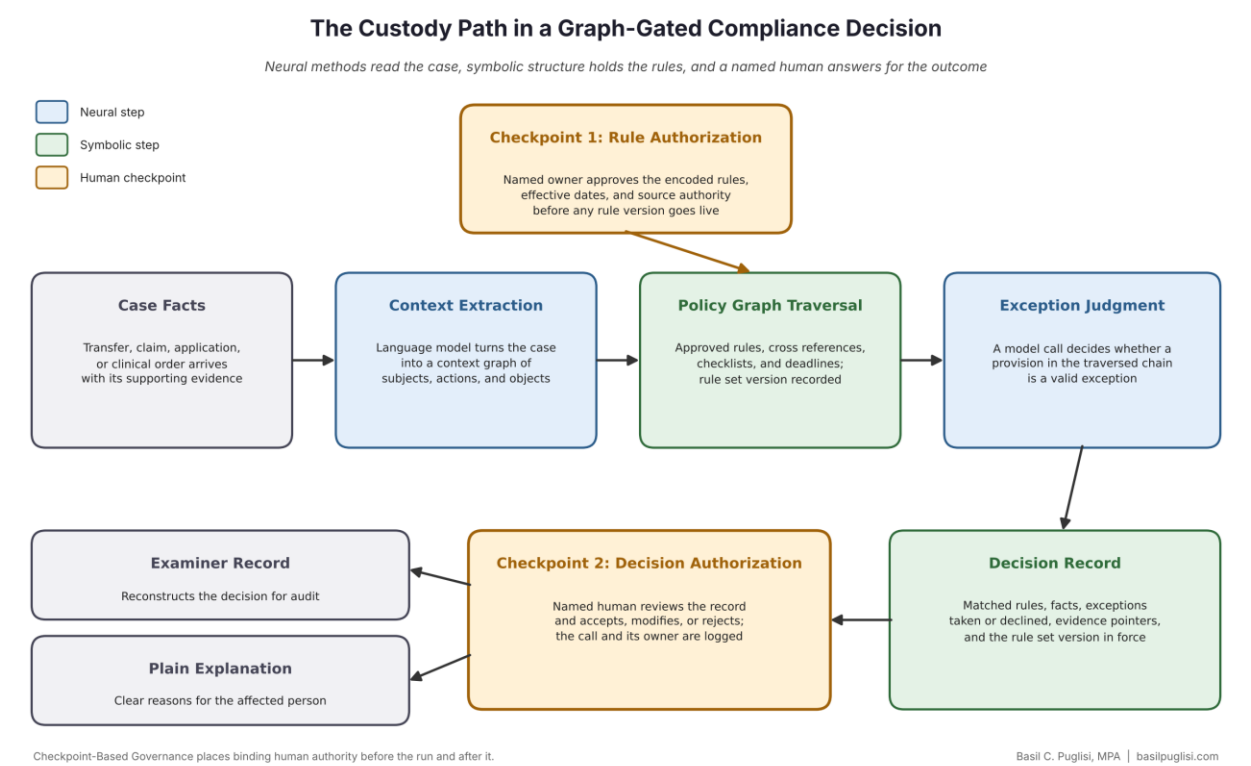
That distinction settles more than it seems to. A compliance system in which a policy graph or a solver checks a language model is still a machine checking a machine, however clean its trail. That is Responsible AI, and for reversible, lower-stakes work it can be a legitimate organizational choice. The same system becomes AI Governance only when a named human holds binding authority at a defined checkpoint and can accept, modify, or reject the result.

[Checkpoint-Based Governance](#) places that authority at two points in a compliance path (Puglisi, 2025). Before the system runs, the human governor sets the scope, the criteria, and the constraints, which in a graph-based design includes approving the encoded rules themselves. After it runs, the

governor validates the output against those criteria and accepts, modifies, or rejects it. The first checkpoint governs what the system is allowed to treat as the rule, and the second governs what leaves the building.

A name in the approval field does not meet that standard. The author's term for the counterfeit is the liability sponge, "a person stationed near an AI decision who lacks the time, the authority, the information, or the standing to actually decide" (Puglisi, 2026). A reviewer who ratifies whatever the system produces is a rubber stamp, and the output stays Responsible AI no matter how senior the title on the approval. The EU AI Act points the same way for high-risk systems, asking that the people assigned to oversight be able to "disregard, override or reverse the output" (Regulation (EU) 2024/1689, Art. 14(4)(d)).

Graphs and solvers are one route to a traversable trail, not the only one. Deterministic rule engines and decision tables already serve that role across many regulated operations, and a language model with disciplined evidence capture can produce a reconstructable record as well. The test is the trail and the authority behind it, and the architecture is a choice made against that test.



What does the law already ask for?

The law already asks for the trail, and it names two audiences for it. Under the General Data Protection Regulation (GDPR), a person has the right not to be subject to a decision based solely on automated processing that produces legal or similarly significant effects. That right has exceptions. It does not apply where the decision is necessary for a contract, authorized by law that also lays down suitable safeguards, or based on explicit consent. In the contract and consent cases, the controller must still offer at least human intervention, a chance to express a point of view, and a way to contest the decision (Regulation (EU) 2016/679, Art. 22). The same person has a right to meaningful information about the logic involved (Regulation (EU) 2016/679, Art. 15). In February 2025 the Court of Justice of the European Union held that this information must describe the procedure and principles actually applied. Neither a complex mathematical formula nor a detailed description of every step of the automated process satisfies that duty (Case C-203/22, 2025).

The EU AI Act requires high-risk systems to allow the automatic recording of events and to be designed so that natural persons can oversee them effectively. It also gives an affected person the right to clear and meaningful explanations of the role the system played in a decision. That right covers decisions based on the output of high-risk systems listed in Annex III, except those in point 2 of that annex, where the decision produces legal or similarly significant effects (Regulation (EU) 2024/1689, Arts. 12, 14, and 86). In United States credit decisions, Regulation B requires a statement of the principal reasons for adverse action and treats a statement that an applicant failed to reach a qualifying score as insufficient (Regulation B, 12 C.F.R. § 1002.9).

None of these laws prescribes a single custody-record format, and none requires the two checkpoints this article describes. Together they set a floor that asks what influenced a decision and who can examine it, and Checkpoint-Based Governance sets an operating standard above that floor.

An examiner needs a record that reconstructs the decision, and the affected person needs a plain explanation of it. A graph traversal can serve the first and still fail the second, so the decision record and the explanation are separate deliverables.

The record that answers the examiner needs only a few fields.

Field	What the record holds
Rule	The obligation that applied, by identifier
Rule version	The version in force on the decision date, with its effective date and source authority
Matched facts	The case facts the rule matched, each with its source

Field	What the record holds
Exception	Each exception considered, and whether it was taken or declined
Evidence	Pointers to the documents behind each fact
Decision status	Pending, accepted, modified, or rejected
Human governor	The named person who holds authority at the checkpoint
Authorization	Whether that person released, modified, or withheld the decision, and when

The technical work can finish while the decision status still reads pending, and that is the point where the record stops and the governor starts.

Why does explainability theater fail high-stakes decisions?

Cynthia Rudin's widely cited perspective in *Nature Machine Intelligence* argues that high-stakes decisions call for models that are interpretable in the first place, not for explanations attached to opaque models after the fact (Rudin, 2019). A post-hoc story about why a black box fired is not the same thing as a decision procedure an auditor can inspect. Large language models widen that gap at scale, since they retrieve, summarize, and narrate with confidence while remaining weak at preserving the cross-references, exception chains, and mutually exclusive checklists that define real regulations (Chung et al., 2025).

Rudin's argument cuts both ways for this article. A hybrid design in which a language model still renders the verdict over a structured plan does not meet her standard for an interpretable model. The structure around the verdict becomes inspectable, while the verdict step itself stays opaque. That is the honest claim for graph-gated systems: they make the path checkable, and the human at the checkpoint stays responsible for the judgment the path cannot make.

IBM Research describes neuro-symbolic AI as combining the strengths of statistical AI, such as machine learning, with human-like symbolic knowledge and reasoning (IBM Research, n.d.). That framing helps practitioners because it refuses a false choice. Verifiability does not require abandoning neural models, and it does require that no neural model serve as the sole custodian of normative logic.

Teams that paste a policy PDF into a prompt and declare the agent compliance-aware have built retrieval with theater lighting. Factics rejects that move for the same reason it rejects a KPI invented after the quarter closes. A method that cannot leave evidence is not ready for a decision that can injure people, trigger fines, or void a certification.

The practical move starts from what Rudin argues and IBM Research describes. High-stakes settings need inspectable decision procedures, and the research line pairs learning with symbolic knowledge and reasoning instead of treating fluent explanation as proof (Rudin, 2019; IBM Research, n.d.). An organization can inventory every AI use case that touches regulated treatment, customer rights, or legal obligation, and mark whether each one produces a rule-level trail or only a narrative rationale. Two measures, both set before the next control review, then do different jobs. The first is the share of those use cases whose stored records let an independent reviewer reconstruct the applied rule chain, sampled each quarter, with a target of 100 percent. The second counts consequential actions released without the authorization the workflow requires, with a target of zero, because a system can extract, match, and validate on its own without holding the authority to release the result.

What does GraphCompliance report about policy graphs versus retrieval?

GraphCompliance, a framework for LLM-based regulatory compliance tested on scenarios derived from the GDPR, makes the architectural bet concrete. It represents regulatory text as a Policy Graph and runtime situations as a Context Graph, then aligns the two before a judge model reasons (Chung et al., 2025). The policy graph encodes normative structure and cross-references, and the context graph formalizes events as subject-action-object and entity-relation triples. The authors report that this alignment reduces the burden of regulatory interpretation and event parsing, which lets the model focus on the core reasoning step.

The paper's diagnosis of failure should sound familiar to anyone who has watched a retrieval pipeline miss a clause. Cross-references dispersed across articles and recitals drop out when retrieval favors query relevance over the links between chunks. Provisions structured as decision trees lose their order-dependent branching, and checklist obligations get merged with deadlines omitted (Chung et al., 2025). GraphCompliance treats those as structural problems rather than as prompts to be worded harder. In its design, graph traversal gathers the chain of cross-referenced provisions, and a second model call then decides whether any provision in that chain is a valid exception, so the structure is deterministic while the exception judgment is not.

On 300 GDPR-derived scenarios across five evaluation tasks, the authors report micro-F1 scores 4.1 to 7.2 percentage points higher than LLM-only and retrieval-augmented baselines (Chung et al., 2025). They also report fewer under- and over-predictions, higher recall, and lower false positive rates. One comparison in the same study sharpens the point. GraphRAG, which uses a graph to improve retrieval, did not yield a meaningful uplift over standard retrieval on this task, while the graph used as a gate outperformed GraphRAG by up to 12.8 percentage points of macro-F1. The best configuration still reaches a micro-F1 near 57 percent, so the reported gain moves a hard problem forward without solving it. The study is a preprint first posted in October 2025, and the

authors withheld the full benchmark for research ethics reasons while releasing its construction protocol and a set of illustrative samples.

Knowledge graphs are not new. Hogan and colleagues document graph-based data models and query languages that represent entities and relationships, along with the deductive and inductive techniques used to reason over them (Hogan et al., 2021). What is new for operators is the pressure to put those graphs in the path of generative systems already running in production. Seen that way, neuro-symbolic compliance is less a research brand than a custody pattern, in which neural methods extract and interpret, symbolic structure constrains and verifies, and provenance stays intact for the human checkpoint.

For a single live workflow, that result points to a concrete step. The study reports that aligning policy and context graphs, with structural checks ahead of the model's judgment, improves GDPR compliance adjudication against LLM-only and retrieval-augmented baselines (Chung et al., 2025). Encoding the governing obligations as an explicit policy graph with cross-reference edges stops the model from rediscovering article structure on every call. Whether that works locally shows in two numbers taken four weeks before and four weeks after the graph gate goes live: the false-negative rate on known violation scenarios in a held-out challenge set labeled by counsel, and reviewer time per case.

Where do solvers and knowledge graphs beat trusting the completion?

Beyond graph-aligned judges, neuro-symbolic compliance work is also bringing formal solvers into the loop. Hsia, Yu, and Jiang describe a framework that pairs large language models with Satisfiability Modulo Theories (SMT) solvers for financial legal analysis. The paper was presented at the 2025 Alware conference, where it received a Distinguished Paper award (Hsia et al., 2025). The language model interprets statutes and enforcement cases into SMT constraints, and the solver enforces consistency and computes the minimal factual modification needed to restore legality when a violation appears. Evaluated on 87 enforcement cases from Taiwan's Financial Supervisory Commission, the system reached 86.2 percent correctness in generating SMT code. The authors report reasoning efficiency gains of more than 100 times in the solver-backed analysis stages, note that the ground-truth encodings were manually corrected and validated, and describe the results as a preliminary foundation.

The design principle matters more than the jurisdiction. When legality is a constraint satisfaction problem, a solver reports the contradiction in the constraints it receives, while a chat model can fail charmingly. The qualification is the encoding. A solver checks only what it was given, and roughly

one in seven generated encodings in that study was wrong before correction. The person who approves the encoded rules therefore matters as much as the solver that runs them.

Healthcare offers a parallel pattern. Bajracharya and colleagues at Datum.md describe the Knowledge Augmented Graph Reasoner, which combines a guideline-based reasoning graph, a benchmarking framework, and a domain knowledge graph to improve instruction adherence in clinical AI (Bajracharya et al., 2025). They frame guideline adherence as a matter of reliability and safety, note that hallucination rates remain high even in recent models, and report adherence correctness above 95 percent on a pharmacogenomics benchmark they built. The work is a short industry abstract presented at the 2025 Conference on Neurosymbolic Learning and Reasoning, and the benchmark is the authors' own, so the figure reads as a working result that awaits outside testing. Across these domains the architecture repeats, because neural systems read the world, symbolic systems retain the rules, and humans retain the signature.

Not every small business needs an SMT stack next week. Regulated teams do need to stop confusing a citation dump with a proof. A compliance agent that cannot show which obligation fired, which facts matched, which exception overrode, and where the graph traversal stopped is not neuro-symbolic governance. It is a better search box.

Choosing among these designs comes down to matching the weight of the symbolic layer to the risk. The published solver-backed and guideline-graph systems treat legal or clinical constraints as executable structure, and their authors report checkable outcomes instead of post-hoc prose (Hsia et al., 2025; Bajracharya et al., 2025). Policy graphs and deterministic gates fit rights and transfer logic, and solvers earn their cost where consistency and minimal repair are the product. Every production decision should then emit a machine-readable decision record, meaning the matched rules, the exceptions, the evidence pointers, and the version of the rule set in force, stored with the case record. Coverage is measured against a target of 100 percent and read beside the error rate counsel finds in a sample of those records, because a record that exists is not a record that is right.

What keeps the trail true after the rules or the facts change?

A policy graph is a legal interpretation frozen into structure, and it ages the way the law does. Regulations change, guidance shifts, and courts read provisions in new ways, so a graph that was right in March can certify a superseded rule in September with perfect traversal. The controls are ordinary change control applied to obligations. Each obligation family needs a named owner, every rule needs an effective date and a source authority, and each rule version needs approval before it goes live. Regression tests against known cases run before release, and a record shows which past

decisions a correction reaches. The approval step is the first checkpoint, and it belongs to someone with the authority to say what the rule means.

Rules are half of the problem. A correct rule applied to facts that are incomplete, outdated, wrongly extracted, or manipulated produces a decision that traverses perfectly and is still wrong. A solver evaluates a mis-encoded fact with the same rigor it gives a sound one, and a governor can approve a record that leaves out what mattered. Evidence integrity therefore needs its own checks, separate from rule integrity: a source for every fact, a sampled independent check of the extraction step, and a record of what the governor saw at the moment of approval. Checkpoint-Based Governance draws the same line, "Authority over what happens is not authority over what is true" (Puglisi, 2025), and Factics starts from the fact for the same reason.

What does the published research disagree about?

Two recent preprints complicate the picture, and both deserve a hearing. Wang and colleagues compared pure language model classification, model-based formal reasoning, and solver-based reasoning on legal entailment in real contracts (Wang et al., 2026). They report that models often presented conclusions as if the solver had produced them without running it, a failure they call scope laundering, and that autoformalization, the step that turns legal text into formal logic, was the primary bottleneck. Sadhu and colleagues deleted, swapped, or negated the governing rule while holding a compliance case fixed, and report that model verdicts were often invariant to those changes (Sadhu et al., 2026).

The first finding cuts against any claim that a solver in the loop settles the question on its own. The second supports the diagnosis that a model can produce a rule-shaped verdict that does not depend on the rule, which is the failure a traversable trail is meant to expose. It also sets a test that any model judge inside a graph-gated design should pass before its verdicts are trusted. Read together, the findings separate three properties that are easy to merge. An output can be correct, the reasoning can be faithful to the procedure it claims to follow, and the decision can be governed, and none of the three implies the others.

What weekly cadence keeps neuro-symbolic work honest?

Measurement defined after the outcome is a narrative, while measurement defined first is a commitment, and a weekly cadence holds that line in practice.

On Monday, pull the week's AI-touched compliance decisions and separate them into trail-complete and narrative-only cases. On Tuesday, for every narrative-only case that would have mattered to a

regulator, write the missing custody fields: facts observed, rules applied, exceptions considered, and any human override. On Wednesday, encode one recurring obligation family into the policy graph or rule store instead of leaving it in a prompt, with the owner's approval recorded before it goes live. On Thursday, run a challenge set that includes cross-references, exceptions, and checklist deadlines, because those are the failure points GraphCompliance documents (Chung et al., 2025). On Friday, review first the errors that cost the most in that workflow. In most compliance work that means false negatives, since a miss usually costs more than an extra human look, but a false positive that blocks a lawful transfer or denies a valid claim carries harm of its own.

What evidence would prove this position wrong?

Suppose an organization runs its own challenge sets for a quarter. Unconstrained language model or retrieval-augmented adjudication then matches graph-gated adjudication on recall, on false positives, and on the agreement rate of a blind re-review sample. It also matches on the measures that test custody directly: independent reconstruction of each decision, the provenance of the evidence, fidelity to the rule version and exceptions in force, and the integrity of the authorization record, all at equal latency and cost. In that case the neuro-symbolic investment weakens for that workflow, and a simpler route to the trail wins. Until local evidence says otherwise, the published structural gains and the audit reality of regulated work favor rules a reviewer can traverse.

How should operators adopt this without boiling the ocean?

Start where "the model said so" is already unacceptable. International transfers, consent and lawful basis, adverse action notices, claims adjudication, clinical guideline deviation, and financial suitability checks are good first rooms, because the cost of narrative failure is already priced into the organization. Use language models for extraction and interface, graphs for obligation structure, and solvers where contradictions must be found and not paraphrased. Keep a named human at the checkpoint for any decision that changes a person's rights or money.

The value here is not another slogan about trustworthy AI. It is an operating rule: if the system cannot show the path from evidence to rule to outcome, it does not belong in an automated compliance path, however high the demo score looks. Showing the path is the floor rather than the ceiling, because where a decision changes a person's rights or money, a named human with the authority to accept, modify, or reject still answers for it. Commodity content says to add guardrails, and this work says what the guardrail is made of.

A first pilot can stay small, because the published hybrid designs add a structural layer around the language model instead of replacing it, so regulated teams can introduce them workflow by

workflow (Chung et al., 2025; Hsia et al., 2025). The right room is a workflow where auditors already reject narrative-only rationales. The pilot carries three deliverables only: a policy graph for one obligation family, a context extraction schema for the cases that reach it, and a reviewer interface that displays the decision record before approval. Success shows first in the median time to evidence for an audit or regulator information request on pilot cases, tracked beside the human override rate. A falling override rate is read with care, since it can mean the graph is right or that reviewers stopped reading, so a sampled blind re-review runs beside it.

What to watch next

Watch whether the compliance AI roadmap still measures success as answer rate and latency instead of record completeness, record correctness, and the false-negative rate on challenge sets. Watch for neuro-symbolic tooling that compiles obligations into graphs or constraints instead of leaving norms in prose (Chung et al., 2025; Hsia et al., 2025). Watch procurement language as well, because a vendor that cannot show how cross-references and exceptions are handled outside the prompt is selling eloquence. Verifiable rules combined with learning, under a named human's authority, are how regulated teams keep AI without surrendering the trail that makes accountability real.

Sources

- Bajracharya, R., Owusu-Obeng, A., Saoulidis, A., Acharya, X., Lo, C. W. H., Bajracharya, A., & Bhandari, D. (2025, September). *Knowledge Augmented Graph Reasoner (KAGR): A neuro-symbolic approach to instruction adherence in healthcare AI* [Industry abstract]. 19th Conference on Neurosymbolic Learning and Reasoning (NeSy 2025), Santa Cruz, CA, United States. <https://2025.nesyconf.org/assets/pdf/papers/nesy2025-KAGR-datum.pdf>
- Case C-203/22, *CK v Magistrat der Stadt Wien and Dun & Bradstreet Austria GmbH*, ECLI:EU:C:2025:117 (Court of Justice of the European Union, February 27, 2025). <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:62022CJ0203>
- Chung, J., Ko, R., Yoo, W., Onizuka, M., Kim, S., Kim, T.-W., & Shin, W.-Y. (2025). *GraphCompliance: Aligning policy and context graphs for LLM-based regulatory compliance* (arXiv:2510.26309) [Preprint]. arXiv. <https://doi.org/10.48550/arXiv.2510.26309>
- Hogan, A., Blomqvist, E., Cochez, M., d'Amato, C., de Melo, G., Gutierrez, C., Kirrane, S., Labra Gayo, J. E., Navigli, R., Neumaier, S., Ngonga Ngomo, A.-C., Polleres, A., Rashid, S. M., Rula, A., Schmelzeisen, L., Sequeda, J., Staab, S., & Zimmermann, A. (2021). Knowledge graphs. *ACM Computing Surveys*, 54(4), Article 71. <https://doi.org/10.1145/3447772>
- Hsia, Y.-S., Yu, F., & Jiang, J.-H. R. (2025, November). *Neuro-symbolic compliance: Integrating LLMs and SMT solvers for automated financial legal analysis* [Conference paper, Distinguished Paper]. 2nd IEEE/ACM International Conference on AI-powered Software (AIware 2025), Seoul, South Korea. <https://arxiv.org/abs/2601.06181>
- IBM Research. (n.d.). *Neuro-symbolic AI*. Retrieved October 11, 2026, from <https://research.ibm.com/topics/neuro-symbolic-ai>
- Puglisi, B. C. (2012, November 27). *Digital Facts: Twitter*. Digital Media Press. <https://www.magcloud.com/browse/issue/471388>
- Current framework page: Puglisi, B. C. (2026). *Facts: The method behind valuable content and trusted AI*. <https://basilpuglisi.com/facts/>
- Puglisi, B. C. (2025, September 23). *Checkpoint-Based Governance*. First public record.
- Current framework page: *Checkpoint-Based Governance: Presence is not authority*. <https://basilpuglisi.com/cbg/>
- Puglisi, B. C. (2026, September 17). *The AI HOAX distraction: Safety, money, and why you can't tell which is driving*. basilpuglisi.com. <https://basilpuglisi.com/ai-hoax/>

Source base: 27 entries, including 3 scholarly articles and papers, 2 preprints, 3 primary documents, 14 organizational, industry, and news sources, and 5 works by the author. Examples: Amodei (2026, September 12), "We must pace the frontier"; Kim, Garg, Peng, and Garg (2025), *Proceedings of the 42nd International Conference on Machine Learning*; *Anthropic PBC v. U.S. Department of War*, No. 3:26-cv-01996 (N.D. Cal., order of August 27, 2026).

Regulation B, 12 C.F.R. § 1002.9 (Notifications). <https://www.ecfr.gov/current/title-12/chapter-X/part-1002/section-1002.9>

Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 (General Data Protection Regulation), Articles 15 and 22. <https://eur-lex.europa.eu/eli/reg/2016/679/oj/eng>

Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 (Artificial Intelligence Act), Articles 12, 14, and 86. <https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng>

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5), 206 to 215. <https://doi.org/10.1038/s42256-019-0048-x>

Sadhu, S., Sengupta, A., Sankarapu, V. K., & Seth, P. (2026). *Verdict without the rule: Diagnosing and auditing regulatory rule sensitivity in LLM compliance systems* (arXiv:2610.12313) [Preprint]. arXiv. <https://arxiv.org/abs/2610.12313>

Wang, O. P., Wong-Toropainen, S., Amrollahi, D., Bai, R., Bansal, T., Garg, A., & Gilpin, L. H. (2026). *Know your limits: On the faithfulness of LLMs as solvers and autoformalizers in legal reasoning* (arXiv:2606.16118) [Preprint]. arXiv. <https://arxiv.org/abs/2606.16118>

Disclaimer

I am not a lawyer, and this article does not provide legal advice. This is thought research and governance analysis based on public sources, cited materials, and human-AI review. It is intended to help executives, practitioners, insurers, and governance teams think more clearly about AI risk, liability exposure, and documentation practices. Readers should not rely on this article as a legal opinion, compliance determination, or substitute for qualified counsel. Any organization facing a legal, regulatory, contractual, or insurance question should consult its own attorney, broker, or professional adviser before acting.

The Other AI: Audio Briefings on Augmented Intelligence and AI Governance

[Spotify](#) | [Apple Podcasts](#) | [Amazon Music](#) | [YouTube Playlist](#)

Basil C. Puglisi, MPA

A Human-AI Collaboration

basilpuglisi.com

#AIassisted using the HAIA Ecosystem | CC BY-NC-SA 4.0

Free for personal, educational, and noncommercial research use with attribution. Commercial exploitation, paid productization, and enterprise commercialization require separate permission and licensing.