Make Law Executable: The Lawie Doctrine

Ten principles for systems that do legal work, derived from the sources of legal authority, professional responsibility and accountability

Position paper LL-DOC-01
Version 1.0.0, 26 November 2026

Abstract

A doctrine states what a system must do and what it must never claim. This paper derives the Lawie doctrine from three bodies of established thought rather than from product ambition: the theory of legal authority, in which a rule's validity is a matter of its pedigree; the law of professional responsibility, in which the lawyer remains answerable for every tool; and the engineering of accountability, in which control is designed into a system and checked at every access. Ten principles follow, beginning with the one that names the suite: intelligence is not authority. Each is stated with its rationale, its consequence for the design of a system, its cost, the evidence that would falsify or weaken it, and where Lawie's own source code stands against it, with an evidence label on every statement about Lawie. The standing is uneven. Some principles are enforced exactly in parts of the code, and several are contradicted by measurements this suite reports. The doctrine is the standard Lawie is measured against, not a description of what it has achieved. It closes with the claims it forbids and the problems it leaves open.

Evidence boundary. Nothing described here is deployed in production [BLOCKED] or has a dated production acceptance [ABSENT]. No component has been independently audited or reviewed by an independent legal expert [ABSENT]. Measurements cited here concern components of the source repository at commit `ffa13d4bce`, executed offline [MEASURED]; the whitepaper LL-WP-01 and the benchmark reports LL-BEN-01 and LL-BEN-02 give the methods. This paper states principles and cites research measurements at a stated code basis. It is not a security assurance, a compliance statement or legal advice, and no independent security, privacy or legal review of it has taken place. Dates of application cited here are research labels from schedules derived from the official texts of the acts; they have not been checked by a qualified lawyer [ABSENT] and are not legal advice.

Prepared by Lawie Labs. *Status:* First edition. *Distribution:* Public. *Basis:* source repository `main @ ffa13d4bce` (2026-11-24); experiments of 26 November 2026; EUR-Lex texts retrieved 26 November 2026. *Audience:* Senior lawyers, general counsel, regulators, legal technologists, security engineers and AI researchers.

Evidence labels. Every material statement about Lawie carries an evidence class. [IMPLEMENTED] in source at the basis commit; [TESTED] exercised by repository tests at that commit; [MEASURED] measured by an offline experiment of this suite on the stated public or synthetic data, never a production or customer measurement; [ANALYTICAL] follows by argument from stated assumptions; [SIMULATED] output of a stated model; [EXTERNAL] checked against a cited primary source; [PROPOSED] design or protocol not built or not run; [UNVALIDATED] a hypothesis, stated as such; [ABSENT] not implemented; [BLOCKED] awaits an external owner or decision. Statements describe the source repository at the basis commit, not a deployed service: no workflow described here has a dated production acceptance, nothing is independently audited, and no statement has been reviewed by an independent human expert. This is a technical publication, not legal advice, not an offer of services and not a statement that any use of Lawie complies with any law.

Contents

Abstract **1**
1 Where authority comes from **3**
1.1 Validity is pedigree 3
1.2 Responsibility stays with a person 3
1.3 Accountability is designed in 3
1.4 What “make law executable” means 4
2 The ten principles **4**
2.1 D1 Intelligence is not authority 4
2.2 D2 Execution, not advice 5
2.3 D3 No anchored source, no decisive claim 6
2.4 D4 We support professionals and render no legal judgment 6
2.5 D5 Law has a clock 7
2.6 D6 An approval binds content, version, time and person 7
2.7 D7 Separation of duties is not optional 8
2.8 D8 Evidence travels with its limitations 9
2.9 D9 Refusal is a result 9
2.10 D10 Measure yourself in public and keep the negative results 10
3 What the doctrine forbids **10**
4 Open problems **11**
References **13**

1 Where authority comes from

1.1 Validity is pedigree

Hart described a legal system as the union of primary rules, which govern conduct, and secondary rules, among them a rule of recognition by which officials identify which rules are valid [1]. Validity under a rule of recognition is a matter of pedigree: a rule is law because it was made in the recognised way and published in the recognised place, not because its content is good. Raz's sources thesis turns the point into a test of method: the existence and content of law are identified by social facts, without resort to moral argument [2].

Pedigree can be checked. Whether a sentence is the text of a given paragraph of a regulation, in the version in force and applicable on a date, published in the Official Journal under a given number, is a question of fact with a mechanical answer once the sources are admitted and their dates transcribed. That is the opening for machines: the law's own test of validity is one that software can apply. It is also the limit. Software can check pedigree; it cannot confer it. A model whose output happens to be correct has not produced law, any more than a well-informed clerk has.

1.2 Responsibility stays with a person

Professional regulation reaches the same place from the other side. The American Bar Association's Formal Opinion 512 applies the existing duties of competence, confidentiality, communication, candour toward the tribunal, supervision and reasonable fees to lawyers' use of generative AI; the duties remain the lawyer's [3]. In its published decision in *Mata v. Avianca*, the court held the lawyers who filed a brief answerable for its citations, whoever drafted them [4]. Article 14 of the AI Act provides for human oversight of high-risk systems; the schedule encoded from Article 113, as amended by Regulation (EU) 2026/1744, lists its application from 2 December 2027 for Annex III systems and from 2 August 2028 for Annex I systems. The article asks that the persons who exercise oversight be enabled, among other things,

> "(b) to remain aware of the possible tendency of automatically relying or over-relying on the output produced by a high-risk AI system (automation bias), in particular for high-risk AI systems used to provide information or recommendations for decisions to be taken by natural persons;"

(Article 14(4), unamended by Regulation (EU) 2026/1744) [5, 6]. A machine can carry much of the work. It cannot carry the responsibility, because responsibility attaches to someone who can be asked to answer.

1.3 Accountability is designed in

Accountability is not a document written after the fact. Raji et al. [7] argue that it must be built through an audit process that runs through the whole development life cycle and leaves documentary artefacts at every stage; management frameworks such as the NIST AI Risk Management Framework and ISO/IEC 42001 ask organisations to assign roles and responsibilities for AI risk and to measure it [8, 9]. Security engineering supplied the mechanisms long before: base access decisions on permission rather than exclusion (fail-safe defaults), check every access against current authority (complete mediation), require two keys where one could do harm (separation of privilege), and grant no more authority than a task requires (least privilege) [10]. Commercial integrity rests on well-formed transactions and the separation of duties [11]. Systems that act with growing autonomy need these mechanisms more, because their harms are more direct and the accountability for them more diffuse [12].

1.4 What "make law executable" means

Lessig [13] observed that code regulates conduct as effectively as law. For a system that executes legal work the observation is literal: a requirement that is not enforced in code is not a property of the system, whatever its policies say. "Make law executable" therefore has two meanings, and the doctrine needs both. In the first, the law's own tests (pedigree, applicability on a date, the mandate of the person who acts) become checks that the system cannot route around. In the second, legal work becomes execution under those checks: a completed act carrying its authority and its evidence, rather than advice left for a human to verify and carry out. Neither meaning is that a machine decides what the law requires. The principles below are built to hold that line.

2 The ten principles

Table 1 lists the principles and the price of each. Every principle is then stated in the same six parts: the statement, its rationale, its consequence for design, its cost, what would falsify or weaken it, and where Lawie stands. The consequence for design describes what a system that holds the principle does; whether Lawie does it is stated only under "where Lawie stands". In that last part every statement about Lawie carries the evidence class that the suite's claim register assigns to it, and "against the principle" introduces findings that contradict it.

Table 1: The Lawie doctrine and what each principle costs.

	Principle	**Cost**
D1	Intelligence is not authority	Machinery narrower and slower than the model it constrains
D2	Execution, not advice	Answerability for every path by which the system acts
D3	No anchored source, no decisive claim	Withholding whatever cannot yet be anchored
D4	We support professionals and render no legal judgment	No substitute for a lawyer to sell
D5	Law has a clock	Per-provision schedules, re-verified on every amendment
D6	An approval binds content, version, time and person	A new approval for every change
D7	Separation of duties is not optional	Enough qualified people for every path
D8	Evidence travels with its limitations	Artifacts that persuade less
D9	Refusal is a result	Lower coverage, and refusals that compound
D10	Measure yourself in public and keep the negative results	Failures published beside successes

2.1 D1 Intelligence is not authority

Statement. A model's competence never confers the right to act or the warrant to be believed.

Rationale. Validity is pedigree and the right to act is conferred (section 1). Neither is a property of an output, so neither improves as outputs improve. The whitepaper LL-WP-01 makes the point formal: the facts on which authorization depends (receipts, identities, times) are not functions of a run's content, so every procedure that decides authorization from content alone errs on some history. Calibrated language models must, in addition, hallucinate on facts that their training data cannot settle [14]; whatever limits that error lies outside the model.

Consequence for design. No model output, whether an answer, a score, a confidence value, an entailment verdict or a model's description of its own authority, is ever an input to a decision about authority. Authority

predicates read records of other people's acts: source admissions, span hashes, approval receipts, identities and times.

Cost. The machinery that holds authority is narrower and slower than the model it constrains. A capability the model has cannot be offered until the machinery covers it, and some of what a model can do will never be authorized.

What would falsify or weaken it. The principle is a claim about kinds and does not weaken as models improve. It would lose practical force if the institutions that test legal work, such as courts, regulators, auditors and counterparties, came to accept a vendor's attestation of model quality in place of sources and mandates. It would be refuted if a model could produce by itself the facts on which authority depends; the principle predicts that it cannot, because those facts are records of what other people did.

Where Lawie stands. Lawie's citation-authority predicate takes no model output as an input [TESTED]. A tool call's authority is the most restrictive of its capability modes [TESTED], and tenant tool grants are default-deny [TESTED]. Against the principle: no component produces a calibrated probability of correctness [ABSENT], and on the LLRB-1 test split the verification pipeline reported the same confidence, 0.973 and HIGH, for supported and unsupported statements alike [MEASURED]. Shown to a user, it presents a property of the pipeline as a property of the answer.

2.2 D2 Execution, not advice

Statement. The product of legal work is a completed act, carried out under recorded authority and accompanied by its evidence, not a suggestion left for a human to verify and perform.

Rationale. Legal consequences attach to acts: the filing, the notice, the signature, the publication. Advice leaves the whole burden of checking and acting with the professional, so the error is relocated, not reduced. A system that executes must answer, for every act, the three questions a hostile reader asks: on what source, by whose authority, and can both be checked without trusting the operator?

Consequence for design. The unit of design is the authority chain: principal, named human authority, approval receipt bound to the exact content, action, artifact. Approval, execution and finality are distinct states, never blurred, and every act leaves an artifact that carries its chain.

Cost. Execution makes the system answerable for every path by which it acts. Each tool needs mediation, receipts and evidence before it is offered, so the product is narrower than an assistant that only advises. Systems that act directly widen both the reach of their errors and the gaps in accountability [12], and error compounds over steps: twenty steps that are each 99% correct leave about one run in five wrong somewhere (LL-WP-01, Proposition 2.2).

What would falsify or weaken it. If supervised execution did not reduce the professional's review burden relative to advice, or if the errors of executed acts cost more than the review saved, the practical case for the principle would fail. Neither can be measured for Lawie until a workflow is in production use, and no Lawie workflow has a dated production acceptance [ABSENT].

Where Lawie stands. The mission states the principle [IMPLEMENTED]. Lawie's authority model binds every consequential action to a principal, a named human authority, a scoped and expiring approval receipt, the action and its artifact, and keeps approval, execution and finality distinct; it is a design, enforced to different degrees on different surfaces [PROPOSED]. Signed portable dossiers carry the evidence of institutional execution and verify offline against pinned keys [TESTED]. No product or workflow has a dated production acceptance [ABSENT].

2.3 D3 No anchored source, no decisive claim

Statement. A decisive legal proposition is released only if it is anchored to an addressable span of an admitted official source, applicable on the relevant date, that fixes every decisive element of the proposition. A citation string is not an anchor.

Rationale. The earlier line, "no citation, no claim", is satisfied by a fabricated citation. Lawie's own citation-required rule shows it: it blocked every legal conclusion without a citation and none with a fabricated one [MEASURED]. A citation is a claim about a source; an anchor is a checked relation to one. The source is admitted, the span is the source's own bytes, the provision applies on the date, and the provision fixes the actor, the deontic operator, the quantities, the dates and the exceptions that the proposition asserts.

Consequence for design. Verification has three layers, and each must be able to refuse on its own: the pedigree of the span (exact, by hash), applicability on the date (by schedule), and support for the decisive elements (typed, with its error model stated). Similarity to retrieved text is not one of the layers: a statement can share almost every word with a provision and say the opposite.

Cost. Much correct legal writing cannot be anchored today. Paraphrase, synthesis across provisions and reasoning from case law have no exact span, and exact matching rejects even typographic variants. Under this principle such output is withheld or returned to a professional.

What would falsify or weaken it. If anchored verification of paraphrase proved impossible at an acceptable false-reject rate, the principle would force quotation-only outputs or wholesale withholding. The study that would measure decisive errors in generated answers is defined and not run [PROPOSED].

Where Lawie stands. The citation-authority predicate implements the first layer exactly: a provision is authoritative only as an exact byte span of an admitted observation, proven by hash and in force at the citation's date [TESTED]. Only one span derivation is admitted, for Czech statutes, so no EU, Slovak or case-law text can be verified [IMPLEMENTED]; with the derivation that EU texts would need, exact-span authority rejected every LLRB-1 item [MEASURED]. A research answer with any unbacked citation is withheld [TESTED]. Against the principle: Lawie's lexical grounding check accepted all 1,959 decisively edited statements of the LLRB-1 test split [MEASURED]. The citation extractor resolved none of 114 citations written in the Official Journal's own style [MEASURED].

2.4 D4 We support professionals and render no legal judgment

Statement. A system that does legal work supports professionals and renders no legal judgment. Its outputs are drafts, analyses and packages for the sign-off of a named professional; it decides nothing about what the law requires of anyone, and nothing it produces is presented as legal advice.

Rationale. Responsibility attaches to a person who can be asked to answer. The duties of competence and supervision presuppose a professional who exercises judgment; a system that rendered judgment would leave the duty without a holder. The AI Act names the failure the boundary guards against: automation bias, the tendency to rely or over-rely on a system's output (section 1).

Consequence for design. Every consequential output names the professional who signs it off. Interfaces show what the system checked and what it did not. Outputs and marketing never imply autonomous legal conclusions, and where the system cannot check, it says so instead of sounding sure.

Cost. The product cannot be sold as a substitute for a lawyer. Its value has to come from execution and evidence rather than from replacing judgment, and the professional's time remains the scarce input.

What would falsify or weaken it. The boundary is nominal if users rely on outputs without the sign-off the design assumes. Override rates, review times and the error rate of reviewed work would show whether it holds; none can be measured until a workflow has a dated production acceptance, which none has [ABSENT].

Where Lawie stands. A research answer with an unbacked citation is withheld, blocks reliance and requires human review [TESTED]. A dossier manifest is valid only if it states, inside the signed manifest, that verification proves neither substantive truth nor admissibility nor legal correctness, and creates no authority [IMPLEMENTED]. Civic Guard excludes the scoring of persons and automatic public accusation, and a repository check blocks wording such as "AI decides" in its copy [IMPLEMENTED]. No component has been reviewed by an independent legal expert [ABSENT].

2.5 D5 Law has a clock

Statement. Every legal proposition is indexed by the date it is about and by the date at which the law is known. Publication, entry into force, application and repeal are distinct facts of each provision, and any of them can move.

Rationale. A provision can be valid law and not yet apply. The schedules encoded from the acts' final provisions as published in the Official Journal list entry into force on 24 May 2016 and application from 25 May 2018 for the GDPR, and entry into force on 10 December 2024 and application from 11 December 2027 for the Cyber Resilience Act [EXTERNAL]. The law also changes after a system has learned it: in the schedule encoded from Article 113 of the AI Act as amended by Regulation (EU) 2026/1744 (in force, by its final provisions, from 27 July 2026), the application of Chapter III Sections 1–3 moves to 2 December 2027 for Annex III systems and to 2 August 2028 for Annex I systems [EXTERNAL]. Temporal databases distinguish valid time from transaction time [15], and legal informatics distinguishes the times at which a norm is in force, efficacious and applied [16].

Consequence for design. Keep per-provision schedules transcribed with verbatim quotations of the final provisions. Key answers to applicability, not to entry into force. Keep the date a question is about apart from the date the law was known. Refuse where applicability depends on facts the system does not have.

Cost. Schedules are expensive to build and must be re-verified on every amendment, and provisions that apply on facts rather than dates must be refused or modelled.

What would falsify or weaken it. The principle describes the law rather than predicting anything about it. Its cost is lowest where application coincides with entry into force and amendment is rare, and highest in the regulatory acts of the last decade.

Where Lawie stands. The temporal index uses half-open validity intervals [TESTED], but the EUR-Lex parser starts each version at entry into force [IMPLEMENTED], and dates of application, transitional provisions and consolidated versions are not modelled [ABSENT]. An index keyed on entry into force answers "applicable" wrongly on every day between entry into force and application [ANALYTICAL]. Against the principle: in LLRB-1 the index and the authority in-force check did so on 781 of 781 such dates [MEASURED], and over 2016–2028 the conflation contradicts the encoded schedules on 7–20% of provision-days [MEASURED]. Lawie's legislative-diff engine found 82 of the 83 article-level amendments of 2026 but read the postponement of application dates as a new obligation [MEASURED].

2.6 D6 An approval binds content, version, time and person

Statement. An approval authorizes exactly one tool, one version and one input, identified by its hash, within a validity window, by a named reviewer who is not the requester.

Rationale. An approval detached from content authorizes whatever content is substituted later. Complete mediation requires that every use be checked against current authority [10], and a well-formed transaction changes data only through certified procedures [11]. In BEN-02 a reference system that treats an approval as a flag on a tool refused only rejected and pending approvals and allowed every attack on binding and validity (LL-BEN-02).

Consequence for design. Approvals are receipts: tool identity, exact version, hash of the canonical input, requester, reviewer, request and review times, expiry and decision. The evaluator recomputes the hash of the current input at the moment of use, and refuses a stale, substituted or self-reviewed approval.

Cost. Every change to an approved input needs a new approval, approvals expire, and long workflows accumulate review latency.

What would falsify or weaken it. Binding secures the content, not the reviewer's attention. If reviewers approve hashes they have not read, the approval is bound and empty. The principle needs evidence of review (what was shown, for how long, how it differed from the last version), which an approval receipt does not by itself contain.

Where Lawie stands. Lawie implements approvals bound to the exact call [TESTED] and decided by authenticated human reviewers [IMPLEMENTED]; the BEN-02 aggregate is given under D10.

2.7 D7 Separation of duties is not optional

Statement. The person who requests an action does not approve it. Where policy requires k approvals, k distinct people give them. The person who reviews a publication does not publish it.

Rationale. Separation of duty is a basic control against error and fraud in commercial practice [11], and a mechanism that needs two keys is harder to subvert than one that needs one [10]. Article 14(5) of the AI Act applies the idea to one category of high-risk system listed in Annex III (from 2 December 2027 in the schedule encoded from amended Article 113); its text provides that the oversight measures for that category are to ensure that

> "[…] no action or decision is taken by the deployer on the basis of the identification resulting from the system unless that identification has been separately verified and confirmed by at least two natural persons with the necessary competence, training and authority."

(Article 14(5), unamended by Regulation (EU) 2026/1744; the bracket omits the words "in addition,") [5].

Consequence for design. Every receipt names requester and reviewer, and the predicate compares them. A quorum counts distinct persons, not approvals. Delegation never lets a sub-agent act where its parent could not. Separation is checked on each path through the system, because it is a property of paths, not of platforms.

Cost. Separation needs enough qualified people for every path. It adds latency, and it is the control that small teams most want to waive.

What would falsify or weaken it. Separation bounds unilateral error and fraud; it does not prevent collusion. Two people acting together defeat it, and nothing in the principle detects that.

Where Lawie stands. Civic Guard's publication path requires a reviewer other than the investigator, a publisher other than the reviewer and a separate authorization for legal language [IMPLEMENTED]. For agent tool calls, see D6.

2.8 D8 Evidence travels with its limitations

Statement. Every artifact states what it establishes and what it does not. Integrity is not truth, a signature is not admissibility, a screen's trigger is not a finding, and a severity label is not a probability.

Rationale. A hash shows that bytes have not changed since they were hashed, and a signature that a key holder attested to them; neither shows that what the bytes say is true, or that a court would admit them. Provenance models make a declared chain of origin explicit, not true [17]. A hash chain shows internal consistency to its holder; only an independent witness of the head makes it evidence against the holder [ANALYTICAL] [18, 19].

Consequence for design. Limitation statements live inside the artifact, not in a terms-of-service page. Every claim stronger than integrity depends on a party outside the operator's control. Scores are labelled for what they are.

Cost. Artifacts that state their limits persuade less than artifacts that do not, and buyers ask for certainty.

What would falsify or weaken it. If readers ignore the limitations, the principle is met in form and void in practice. We are not aware, in the literature cited here, of a measurement of whether limitation statements change how legal readers rely on an artifact.

Where Lawie stands. A dossier manifest is valid only if it states, inside the signed manifest, that verification proves neither substantive truth nor admissibility nor legal correctness, and creates no authority [IMPLEMENTED]. Dossiers verify offline against pinned keys [TESTED], and they failed every tampering in the evidence-integrity study [MEASURED]. Civic Guard marks its screens as not calibrated [IMPLEMENTED] and treats severity as an editorial label, not a probability [IMPLEMENTED]. The audit-chain verifier detected every tested in-place edit, deletion, reordering and insertion and every deleted prefix [MEASURED].

2.9 D9 Refusal is a result

Statement. When a source is missing, a date unresolved, an approval absent or a verifier blind to the error type, the system refuses and says why. A refusal is a first-class, typed and recorded output, and every degraded mode fails closed.

Rationale. Selective prediction trades coverage for risk [20, 21], and a gate lowers the error of what it releases only as far as its false-accept rate is low. Language models guess because evaluation rewards guessing [22]; a system that must not guess must be built, and scored, to prefer an honest refusal to a confident error. Fail-safe defaults are the security form of the same rule [10].

Consequence for design. Refusal reasons are typed and logged. Degraded modes refuse rather than release. Release is reported beside refusal so that the trade is visible. Missing evidence leads to a refusal, not to a softer answer.

Cost. Coverage falls, users wait for professionals, and refusals compound over workflows: with a false-reject rate of 5% and no errors at all, a twenty-step gated run halts somewhere with probability of about 0.64 (LL-WP-01, Proposition 2.2).

What would falsify or weaken it. If refusals were so frequent that users routed around the system, refusal would move risk outside it, where nobody sees it. The principle needs measured refusal rates and their reasons.

Where Lawie stands. The citation-authority predicate refuses with a typed reason for each failed condition, from an unapproved source contract to a version not in force on the cited date [TESTED]. Research answers with unbacked citations are withheld [TESTED]; the citation gate withholds outputs whose metrics are

missing or failing [TESTED]; Civic Guard's screens can abstain [IMPLEMENTED]; and the orchestrator executes nothing when it lacks its verification key or replay store [TESTED]. Against the principle: case-law good-law checks cannot pass at all, because the validity service returns no determination, which is a missing capability rather than a considered refusal [IMPLEMENTED]; outside production the verification pipeline defaults to a degraded-review policy [IMPLEMENTED], under which it released with HIGH confidence a claim that had no retrieved source but carried a citation string [MEASURED]; and the missing-document detector reports absent sources without refusing [TESTED].

2.10 D10 Measure yourself in public and keep the negative results

Statement. Claims about the reliability of a legal system are made only from measurements whose hypotheses were fixed in advance, whose labels do not come from the system measured, and whose negative results are published with the positive.

Rationale. Pre-registration separates prediction from postdiction [23]; benchmark results are only as valid as the benchmark [24]; and reporting standards for machine-learning-based science ask for the disclosures that make work checkable [25]. For legal research tools, Magesh et al. [26] report measured hallucination rates that contradicted the providers' claims. A field whose claims outrun its measurements needs a standard that a competitor or a regulator can apply.

Consequence for design. Every statement about the system carries an evidence class bound to a register. Benchmarks are pre-registered, labelled by construction or by independent annotators, and run on unchanged components whose hashes are recorded. Failures are reported in the same table as successes.

Cost. Publishing negative results exposes weaknesses to competitors and buyers, claims arrive more slowly than marketing would like, and the discipline takes engineering time that ships no feature.

What would falsify or weaken it. Self-measurement carries a conflict of interest that pre-registration narrows and does not remove; without independent replication the principle is a promise. It fails silently if an organisation stops publishing its failures, which is why every statement here carries an identifier that the internal claim register resolves.

Where Lawie stands. LLRB-1's hypotheses, thresholds and test-split checksums were committed before its single confirmatory run [MEASURED]; the development split was explored first, and the registration lists no deviation [MEASURED]; each run records the commit and the hashes of the component files it loads [MEASURED]. The company's claims policy defines truth states and prohibits unqualified claims such as "zero hallucinations" [IMPLEMENTED]. Across the 552 BEN-02 scenarios Lawie's evaluated functions denied none of the 96 legitimate requests, refused every scenario of some attack classes and allowed every scenario of others; the class-level outcomes are recorded in restricted engineering records pending remediation [MEASURED]. Not yet meeting the principle: Lawie's evaluation laboratory has never run a model, and its datasets named after public benchmarks are templated fixtures [ABSENT]; no lawyer has reviewed the benchmark labels [ABSENT]; and no independent audit or expert review exists [ABSENT].

3 What the doctrine forbids

The company's claims policy prohibits a list of unqualified claims [IMPLEMENTED]. The doctrine restates them as commitments, each with the principle it follows from.

Production. We do not call anything production ready because code exists. Nothing described in this suite is deployed in production [BLOCKED] or has a dated production acceptance [ABSENT] (D10).

Error. We do not claim zero hallucinations, or that any output is free of error. We publish the measured error rates of components and say what has not been measured: the study of generated answers is defined and not run [PROPOSED] (D3, D10).

Coverage. We do not claim to cover all law. One span derivation, for Czech statutes, is admitted [IMPLEMENTED] (D3).

Judgment. We do not describe Lawie as an autonomous lawyer, and we do not say that AI decides (D1, D4).

Compliance. We do not say that Lawie, or any use of it, complies with the AI Act or any other law. Compliance is a property of a use in its context, which this paper does not assess (D4).

Tamper evidence. We do not call records tamper-proof. A hash chain is tamper-evident at most, and against its keeper only if an independent party witnesses its head [ANALYTICAL] (D8).

Truth and admissibility. We do not say that verification proves truth, or that any artifact is court-admissible by design; Lawie's own dossier manifests state the opposite [IMPLEMENTED] (D8).

Data. We do not say that Lawie cannot access customer data or that it retains nothing; this suite measures neither (D8, D10).

Classification and attribution. We do not call data classified because a field says so, and we do not claim to detect foreign agents or traitors (D4).

Civic findings. We do not call Civic Guard material verified evidence, and we call no person or institution guilty or an offender; Civic Guard excludes the scoring of persons and automatic public accusation [IMPLEMENTED] (D4, D8).

Superlatives. We do not call anything unbeatable or guaranteed (D10).

4 Open problems

The doctrine leaves problems that, to our knowledge, no one has solved in general. They are stated so that progress on them can be recognised.

1. *Support for paraphrase.* We know of no verifier that decides support for unrestricted paraphrase with a known false-accept rate on decisive errors. In LLRB-1, canonical matching was exact on quotations and their typographic variants [MEASURED]; no item tested paraphrase.
2. *Completeness of error types.* Typed verification protects only the element types it checks; a prototype that typed six kinds of decisive element accepted 88.5% of edits of three kinds it did not type [MEASURED]. No typology of decisive errors is known to be complete for any legal domain.
3. *Base rates.* The error of anything a gate releases depends on the base rate and the type mix of errors in the answers it is given. For generated legal answers in Lawie's pipeline both are unmeasured [PROPOSED].
4. *Provenance.* How to carry a request's provenance through a tool-using agent, without trusting the agent, so that an authorization layer can act on it, is open [27, 28].
5. *Witnesses.* Who witnesses the head of an evidence chain, independently of the operator and at a cost a legal workflow can bear, is an institutional question as much as a technical one.
6. *Quality of review.* Binding an approval secures its content, not the reviewer's attention. How to establish that a named reviewer actually reviewed is open.
7. *Calibration.* No Lawie component produces a calibrated probability of correctness [ABSENT]. Whether any legal verifier can be calibrated for a stated use, and on what population, is open [29, 30].
8. *Applicability that depends on facts.* Some provisions apply when a product is placed on the market or a contract is concluded, not on a date. A system must model the facts or refuse, and where the line between the two should fall is not settled.

A doctrine is worth what it costs to keep. Each principle above has a price, stated with it; Lawie pays some of those prices today and not others, and the measurements and negative results of this suite are reported, with claim identifiers that the internal register resolves, so that anyone can check which.

References

[1] H. L. A. Hart. *The Concept of Law*. Clarendon Press, Oxford, 1961.

[2] Joseph Raz. *The Authority of Law: Essays on Law and Morality*. Clarendon Press, Oxford, 1979.

[3] American Bar Association Standing Committee on Ethics and Professional Responsibility. Formal Opinion 512: Generative Artificial Intelligence Tools, 2024. URL https://www.americanbar.org/content/dam/aba/administrative/professional_responsibility/ethics-opinions/aba-formal-opinion-512.pdf. 29 July 2024.

[4] United States District Court for the Southern District of New York. Mata v. Avianca, Inc., No. 22-cv-1461 (PKC), Opinion and Order on Sanctions, 2023. URL https://www.courtlistener.com/docket/63107798/mata-v-avianca-inc/. 22 June 2023.

[5] European Parliament and Council of the European Union. Regulation (EU) 2024/1689 of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). Official Journal of the European Union, L, 2024/1689, 12.7.2024, 2024. URL https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32024R1689.

[6] Publications Office of the European Union. Regulation (EU) 2024/1689, consolidated text of 27 July 2026. EUR-Lex, document 02024R1689-20260727, 2026. URL https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:02024R1689-20260727.

[7] Inioluwa Deborah Raji, Andrew Smart, Rebecca N. White, Margaret Mitchell, Timnit Gebru, Ben Hutchinson, Jamila Smith-Loud, Daniel Theron, and Parker Barnes. Closing the AI Accountability Gap: Defining an End-to-End Framework for Internal Algorithmic Auditing. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 33–44, 2020. doi: 10.1145/3351095.3372873.

[8] National Institute of Standards and Technology. Artificial Intelligence Risk Management Framework (AI RMF 1.0). Technical Report NIST AI 100-1, U.S. Department of Commerce, 2023.

[9] ISO/IEC. ISO/IEC 42001:2023 Information technology — Artificial intelligence — Management system. International Organization for Standardization, 2023. URL https://www.iso.org/standard/42001.

[10] Jerome H. Saltzer and Michael D. Schroeder. The Protection of Information in Computer Systems. *Proceedings of the IEEE*, 63(9):1278–1308, 1975. doi: 10.1109/PROC.1975.9939.

[11] David D. Clark and David R. Wilson. A Comparison of Commercial and Military Computer Security Policies. In *Proceedings of the 1987 IEEE Symposium on Security and Privacy*, pages 184–194, 1987. doi: 10.1109/SP.1987.10001.

[12] Alan Chan, Rebecca Salganik, Alva Markelius, Chris Pang, Nitarshan Rajkumar, et al. Harms from Increasingly Agentic Algorithmic Systems. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, pages 651–666, 2023. doi: 10.1145/3593013.3594033.

[13] Lawrence Lessig. *Code and Other Laws of Cyberspace*. Basic Books, New York, 1999.

[14] Adam Tauman Kalai and Santosh S. Vempala. Calibrated Language Models Must Hallucinate. In *Proceedings of the 56th Annual ACM Symposium on Theory of Computing*, pages 160–171, 2024. doi: 10.1145/3618260.3649777.

[15] Christian S. Jensen, Curtis E. Dyreson, et al. The Consensus Glossary of Temporal Database Concepts — February 1998 Version. In *Temporal Databases: Research and Practice*, volume 1399 of *Lecture Notes in Computer Science*, pages 367–405. Springer, 1998. doi: 10.1007/BFb0053710.

[16] Monica Palmirani, Guido Governatori, and Giuseppe Contissa. Modelling Temporal Legal Rules. In *Proceedings of the 13th International Conference on Artificial Intelligence and Law*, pages 131–135, 2011. doi: 10.1145/2018358.2018378.

[17] Luc Moreau and Paolo Missier. PROV-DM: The PROV Data Model. W3C Recommendation, 2013. URL https://www.w3.org/TR/prov-dm/.

[18] Stuart Haber and W. Scott Stornetta. How to Time-Stamp a Digital Document. *Journal of Cryptology*, 3(2): 99–111, 1991. doi: 10.1007/BF00196791.

[19] Scott A. Crosby and Dan S. Wallach. Efficient Data Structures for Tamper-Evident Logging. In *Proceedings of the 18th USENIX Security Symposium*, pages 317–334, 2009. URL https://www.usenix.org/conference/usenixsecurity09/technical-sessions/presentation/efficient-data-structures-tamper-evident.

[20] C. K. Chow. On Optimum Recognition Error and Reject Tradeoff. *IEEE Transactions on Information Theory*, 16(1):41–46, 1970. doi: 10.1109/TIT.1970.1054406.

[21] Yonatan Geifman and Ran El-Yaniv. Selective Classification for Deep Neural Networks. In *Advances in Neural Information Processing Systems 30*, pages 4878–4887, 2017.

[22] Adam Tauman Kalai, Ofir Nachum, Santosh S. Vempala, and Edwin Zhang. Why Language Models Hallucinate, 2025. arXiv:2509.04664.

[23] Brian A. Nosek, Charles R. Ebersole, Alexander C. DeHaven, and David T. Mellor. The Preregistration Revolution. *Proceedings of the National Academy of Sciences*, 115(11):2600–2606, 2018. doi: 10.1073/pnas.1708274114.

[24] Samuel R. Bowman and George E. Dahl. What Will it Take to Fix Benchmarking in Natural Language Understanding? In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4843–4855, 2021. doi: 10.18653/v1/2021.naacl-main.385.

[25] Sayash Kapoor, Emily M. Cantrell, Kenny Peng, Thanh Hien Pham, Christopher A. Bail, et al. REFORMS: Consensus-based Recommendations for Machine-learning-based Science. *Science Advances*, 10(18): eadk3452, 2024. doi: 10.1126/sciadv.adk3452.

[26] Varun Magesh, Faiz Surani, Matthew Dahl, Mirac Suzgun, Christopher D. Manning, and Daniel E. Ho. Hallucination-Free? Assessing the Reliability of Leading AI Legal Research Tools. *Journal of Empirical Legal Studies*, 2025. doi: 10.1111/jels.12413.

[27] Kai Greshake, Sahar Abdelnabi, Shailesh Mishra, Christoph Endres, Thorsten Holz, and Mario Fritz. Not What You've Signed Up For: Compromising Real-World LLM-Integrated Applications with Indirect Prompt Injection. In *Proceedings of the 16th ACM Workshop on Artificial Intelligence and Security*, pages 79–90, 2023. doi: 10.1145/3605764.3623985.

[28] Edoardo Debenedetti, Jie Zhang, Mislav Balunović, Luca Beurer-Kellner, Marc Fischer, and Florian Tramèr. AgentDojo: A Dynamic Environment to Evaluate Prompt Injection Attacks and Defenses for LLM Agents. In *Advances in Neural Information Processing Systems 37 (Datasets and Benchmarks Track)*, 2024.

[29] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On Calibration of Modern Neural Networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1321–1330, 2017.

[30] Anastasios N. Angelopoulos and Stephen Bates. Conformal Prediction: A Gentle Introduction. *Foundations and Trends in Machine Learning*, 16(4):494–591, 2023. doi: 10.1561/2200000101.