Outside doctrine, before doctrine

Why agentic healthcare will first be regulated by insurers, contracts, and courts

1. Four regulatory events, one structural negative act

Between January and April 2026, four regulatory acts were produced whose serial reading has not yet received common interpretation.

1. On 12 March 2026, the Centers for Disease Control and Prevention publishes *Considerations for Agentic Research in Public Health*, a US federal document explicitly distinguishing the agentic from conversational generativity and setting out operational principles for its use in public health. The document is followed the next day by a *Considerations for Generative AI in Public Health* that, by contrast, marks its specificity.
2. On 2 April 2026, the Food and Drug Administration addresses to Purolea Cosmetics Lab, an OTC homeopathic manufacturer in Livonia (Michigan) operating under CGMP pursuant to 21 CFR parts 210 and 211, a warning letter containing a section titled *Inappropriate Use of Artificial Intelligence in Pharmaceutical Manufacturing*. The grievance is filed under 21 CFR 211.22(c): the Quality Unit had delegated to AI agents the drafting of specifications, procedures, and master production records without competent review. The firm has since ceased its pharmaceutical production.
3. In January 2026, Mindgard discloses to Doctronic, a clinical AI assistant deployed in a regulatory sandbox in Utah, a chain of vulnerabilities exploitable through context injection. Public report publication follows in March 2026: an attacker can have the agent ingest a false regulatory bulletin, modify its recommendations, and durably contaminate the patient context through the SOAP notes generated. A tripled OxyContin dose can be routed to a human clinician for validation, presented as a structured clinical document.
4. On 2 February 2026, the European Commission misses for the second time the statutory deadline set by Article 6 §5 of the AI Act: the guidelines on high-risk system classification are not published. The Digital Omnibus proposal of 19 November 2025 provides for aligning the application of high-risk rules with the effective availability of harmonized CEN-CENELEC standards, whose first deadline had itself been missed in autumn 2025.

To these four positive acts is added a structural negative act: the European Commission withdrew, in February 2025, its proposed AI Liability Directive. The civil liability framework specific to AI, which could have qualified distributed emergent behaviors, is abandoned. The revised Product Liability Directive 2024/2853 remains; its scope and limits are discussed in §6.

Taken individually, none of these events establishes a stabilized regulatory doctrine. Taken in sequence, they reveal a fact that usual commentary misses: regulatory doctrine is not moving toward risk; it is being bypassed by other forces (economic, contractual, jurisprudential) that decide in its place. The present article argues that this substitution is underway, that it is structural, and that it will determine the effective regulation of agentic healthcare for years to come.

2. Three loci, not two

Usual commentary reads regulatory inadequacy as a two-term dissociation: what regulation names, and what systems do. This reading implicitly fuses behavior production and the risk it carries. It is too simple to grasp a system that does not forgive simplifications.

Three distinct loci must be named.

1. *The locus of behavior production*: orchestration. This is where the effective behavior of the system is composed, at execution, from models, rules, memories, and data. It is not reducible to a component.
2. *The locus of risk*: distributed emergence. This is where a behavior misaligned with the regulated finality arises. This locus does not coincide with the previous one: some orchestrations produce behavior without producing critical risk; some risks appear outside orchestration (through data drift, through defective human integration, through unqualified exposure to external sources). Production is necessary but not sufficient for risk.
3. *The locus of regulation*: the institutional anchor points. This is where regulation can effectively intervene, control, and sanction. It is defined by objects: devices, identifiable operators, auditable processes.

These three loci are not equivalent. Their non-coincidence is the structural fact that no current regulatory framework operates explicitly. Doctrine conflates production and risk, as if governing the first sufficed to contain the second. It also conflates regulation and the locus of risk, as if anchoring them together sufficed to make them coincide. Neither of these two confusions holds under the test of a persistent agentic chain.

3. Conceptual antecedents

The central property defended in this text is not invented ex nihilo. It inscribes itself in a lineage of analyses that have, at different levels, identified the inadequacy of classical imputability frameworks faced with complex systems.

- The theory of *normal accidents* (Perrow, 1984) establishes that in systems with tight coupling and complex interactions, certain accidents are structurally non-attributable to a single cause: they emerge from the combination of individually tolerable failures. This line thematizes the irreducibility of causality in complex systems.
- The *accountability gap*, widely developed since Mittelstadt et al. (2016) in data and AI ethics, designates the gap between the complexity of algorithmic systems and the capacity of legal and organizational frameworks to identify responsible parties. This line thematizes the legal question.
- *Distributed responsibility* (Floridi & Sanders, 2004; developed in subsequent work on artificial moral agency) recognizes that responsibility, in multi-agent systems, may be structurally distributed rather than attributable to a single agent. This line thematizes the ethical question.
- *Causal opacity* and the *traceability gap*, in technical literatures on explainable AI and algorithmic audit, designate the practical impossibility of following the causal chain internal to a model or a system.

These lines cross without strictly coinciding. They share the idea that certain behaviors are not attributable without loss of pertinent information. The present analysis does not claim to refound this idea. It claims to specify it for a particular class of systems, persistent agentic chains, where three characteristics (dynamic composition, structured persistence crossing sessions, exposure to sources external to the certified perimeter) produce a specific, observable, and legally consequential non-localizability.

4. Causal non-localizability: specialization to persistent agentic chains

Causal non-localizability does not designate the absence of causality. It designates the practical and legal impossibility of attributing observable behavior to a single transition without losing pertinent information about the trajectory. It extends the notions of *accountability gap* and *distributed responsibility*, but specifies them in a particular case: persistent agentic chains where memory, external context, and dynamic composition modify the output across sessions.

Stated in its structural form:

For an orchestration chain maintaining structured memory across sessions, observable failure is not, in general, attributable to an identifiable step of the chain without loss of pertinent information about the trajectory. This non-attribution is a property of the chain, not a defect of observation.

Three constitutive characteristics produce it:

1. *Dynamic composition* (model selection depends on the execution context);
2. *Structured persistence* (information transits and reformulates between sessions via typed artifacts);
3. *Exposure to sources external to the certified perimeter* (the inference context is partially unqualified).

These three characteristics are precisely what gives agentic systems their clinical value. They cannot therefore be suppressed to facilitate regulation, nor can one hope for a useful agentic chain that would not manifest them.

The consequence for classical regulatory frameworks is of a structural, not contingent, order. These frameworks (SaMD, AI Act, CGMP) operate under an implicit causal attribution hypothesis: for an operator to be held responsible for a behavior, the behavior must be imputable to an identifiable step. When this hypothesis ceases to hold, the framework no longer captures its object as an autonomous regulatory object; it continues to capture effects and responsibilities through indirect holds, but it ceases to formalize the generator of the behavior.

5. Doctronic: documented technical proof of existence

The Doctronic case is not an illustration of the phenomenon. It is a documented technical proof of existence, in the sense that it establishes, on a system legitimately deployable within a formally valid regulatory framework, that the property stated in the previous paragraph is technically observable. The material comes from a published security report, not from a regulatory act or a jurisdictional decision; this very nature of the material participates in the argument.

The attack does not target the model's weights. It does not target a single session. It exploits the persistence of SOAP notes (structured clinical summaries the assistant generates for the clinician) as a vector of propagation. A false regulatory bulletin injected in a compromised session contaminates the patient context, which is re-injected in subsequent sessions, presented as legitimate clinical context, and ends up validated by a human clinician.

The failure trajectory crosses three moments: the injection window (session N), the transduction of manipulated content into a structured SOAP note (session/persistence

boundary), and confident human reading (session N + k). None of these three moments, taken in isolation, suffices to produce the observed behavior. None, consequently, suffices to qualify it legally. The cause distributes within the trajectory, exactly as the specialization stated above predicts.

Three consequences follow.

1. The trust boundary is not inscribed in the system: the SOAP note is treated by the clinician as belonging to a register of authority (that of the medical record), whereas it originates from a boundary where user content has been promoted to regulated context without explicit qualification.
2. Human oversight is technically present and substantially absent: the clinician validates, but what is validated is no longer reconstructible from the output observed.
3. Imputability, finally, is non-assignable without convention: the deployer can invoke the attacker, the attacker is not legally qualified, the validating clinician followed procedure, and the Utah regulatory sandbox does not cover this case. Any attribution will therefore be *decided*, by contractual or jurisprudential convention, not *established* by the regulatory framework.

Doctronic, in the state of publicly available elements, thus establishes that causal non-localizability is not a conceptual abstraction but a technically observable property on systems in operation. The generalization of this observation to all architectures of the same type is an open empirical question; its relevance as a proof of existence is not.

6. AI Act and PLD: solid indirect holds, not explicit formalization

An honest critique must name what European doctrine does better than American doctrine. Four anchor points deserve to be mobilized without underestimation.

- **Article 25** defines distributed responsibility along the value chain: any operator who substantially modifies a high-risk system becomes a provider.
- **Article 14** imposes effective human oversight.
- **Article 15** requires robustness, accuracy, and cybersecurity, including resilience to manipulation.
- **Articles 56 and 95** structure the Codes of Practice and codes of conduct, instruments the Commission has mobilized to address, downstream of doctrine, what it could not formalize upstream.

These holds are solid. They are not the formalization of agentic orchestration as an autonomous regulatory object. The distinction is precise and deserves to be held.

Article 25 addresses static composition: who becomes responsible for what when an identifiable operator act modifies the system. It does not formalize dynamic composition, where behavior modification results from an execution trajectory without a qualifiable operator act. The *substantial modification* it requires is a legal category, not a behavioral one.

Article 14 addresses oversight as a formal requirement: the system must *enable* oversight. It does not formalize oversight as an effective property, that is, the fact that the output be *actually* analyzable by a human placed in ordinary cognitive conditions. Doctronic satisfies Article 14 in its letter: the clinician validates. Doctronic violates it in its function: validation is captive to a presentation that short-circuits its possibility.

Article 15 addresses the robustness of *the system as defined*, that is, the device in the MDR sense or the AI system in the AI Act sense. It does not formalize the robustness of the *execution trajectory* that crosses several systems, several sessions, and several trust boundaries. Article 15 robustness is a property of the certified perimeter. Risk, however, exits the perimeter.

Articles 56 and 95, by their very existence, signal what the preceding articles cannot accommodate: the Commission delegates to GPAI providers the task of structuring applicable requirements, that is, implicitly acknowledges it cannot formalize what it regulates. This acknowledgment is doctrinally important. It is precisely the entry point of the third regulatory regime (§8).

The PLD 2024/2853 deserves the same precision. It effectively modernizes the product liability regime by including software, product-related digital services, updates, and cybersecurity defects among the elements susceptible to constitute a defect. It can therefore capture the defect of a software component (including AI) or of an integrated digital service, taken individually. It does not, however, capture multi-actor, multi-session, multi-boundary trajectorial causality as an autonomous object, that is, the failure that emerges from composition rather than from any of its components. The withdrawal of the AI Liability Directive aggravates this limit: the initial proposal targeted causality presumptions specific to AI, capable of capturing distributed behaviors without single cause. Its disappearance replaces imputability within a regime that knows how to handle components (and their defects), but does not know how to handle trajectories as such.

European doctrine, on the agentic, is not naive. It is partially inadequate, which is not the worst defect.

7. Purolea: cumulative divergence

The Purolea event does not signal a regulatory displacement. It signals the opposite: the stability of regulation, doctrinally maintained, facing a risk that drifts without it.

The FDA did not regulate the agentic. It applied 21 CFR 211.22(c), an existing provision on the Quality Unit's responsibility, to a case where an AI agent had replaced human oversight. The event is doctrinally orthodox: the FDA defends its regulatory perimeter by refusing to let the AI agent displace responsibility. The implicit message is exact: *regulation stays where it is; it is up to the operator to compensate organizationally for what the agent does not deliver in compliance terms*.

This doctrinal stability has a cost not immediately visible. With every cycle of agentic innovation, a new class of distributed behaviors appears in deployed architectures. With every cycle, regulation maintains its perimeter. The surface of effective risk thus moves away, with every cycle, from the regulated perimeter. This is not merely a static asymmetry between the locus of risk and the locus of regulation. It is a **cumulative divergence**: the gap widens with every innovation, and doctrinal catch-up becomes progressively infeasible without explicit overhaul.

Current regulatory doctrine is in the following situation: it can remain stable for a long time, and every day of stability increases the gap between what it governs and what effectively decides.

8. Three regulatory regimes, not two

The binary "regulation follows risk vs. regulation stays where it can hold" is useful but incomplete. A third regime exists, operating, and largely unthematized by doctrine.

1. *First regime: formal extension*. The regulated perimeter is extended to absorb orchestration: SaMD overhaul (MDR Article 2(1), MDCG 2019-11), qualification of the execution trajectory as a certifiable object, dynamic audit infrastructure. This path is conceivable but costly, slow, and politically difficult. It is not, in its current state, supported by the regulatory acts observed.

2. *Second regime: maintenance and functional stratification*. Regulation preserves its anchor point and imposes procedural compensations on operators: mandatory AI literacy (AI Act Article 4), Quality Unit responsibility (CGMP), distributed data governance. Purolea falls under this regime. It is doctrinally conservative and operationally common.

3. *Third regime: industrial self-regulation constrained by ex post liability.* Platforms define and impose their own orchestration rules: model guardrails, tool policies, context filters, internal logging, terms of service contractualizing the perimeter of

use. This regulation is not optional. It is rendered mandatory by the joint pressure of civil liability (PLD 2024/2853), of contractual liability toward institutional deployers, and of reputational liability facing publicly instructed incidents.

This third regime is already largely institutionalized, in a place doctrine has not read correctly: the Codes of Practice and codes of conduct of Articles 56 and 95 of the AI Act itself. When the AI Act delegates to GPAI providers the task of structuring applicable requirements for systemic-risk models, it implicitly recognizes that doctrine cannot formalize what it regulates. The third regime is not a circumvention of the European framework; it is partially an instantiation of it that the framework does not assume as such.

Industrial self-regulation does not wait for doctrine to exist. It is already the doctrine, in zones doctrine does not formalize. It is set to be dominant on agentic orchestration in healthcare before any formal doctrinal evolution.

9. Economics as engine, not as context

The effective regulatory trajectory will not be determined by doctrine. It will be determined by three economic forces that do not wait for doctrine to decide.

1. *Insurers*. Cyber and professional civil liability coverage for operators deploying agentic systems in healthcare is being redefined. The first explicit exclusions for non-auditable AI systems appear in 2025 on the American market. The rewriting of insurable perimeters (differential premiums based on the presence of audit mechanisms, progressive withdrawal of coverage on unqualified architectures) precedes regulation. It *defines* regulation, through structural effect: a non-insurable system is not deployable, regardless of its formal regulatory compliance.
2. *Courts*. Civil liability under PLD 2024/2853 will see its first cases instructed on agentic architectures in healthcare. The question is known: who pays when a patient suffers damage following a recommendation produced by a chain whose cause distributes? It will be ruled *case by case*, by jurisprudence, before any regulatory doctrine has formalized the conditions. The effective shape of liability will emerge from a few structuring decisions, aggregated by precedent effect. These decisions will impose on deployers operational constraints that regulation will then take up, perhaps, through implementation.
3. *Industrials*. Facing these two forces, operators' arbitrages are not between compliance and non-compliance. They are between three positions: strict compliance (which slows time-to-market and increases marginal cost), contractual externalization (which transfers residual risk to the user via terms of service and disclaimers), and preventive self-regulation (which invests in audit substrates to offer contractual guarantees to institutional clients). The third term

is, for actors with reputational exposure, the most rational. It directly feeds the third regime.

This dynamic is not unprecedented. In post-2010 finance, the effective framing of algorithmic trading was largely formed by SEC enforcement actions, industry risk management standards, and FINRA conventions, adopted in reaction to the flash crash of 6 May 2010 and to the 2012 Knight Capital failure, rather than by doctrinal anticipation; Regulation SCI of 2014 codifies *after the fact* practices the actors had already had to impose. In automated aeronautics, FAA certification was reformed by the Aircraft Certification, Safety, and Accountability Act of 2020 *in response* to NTSB investigations and litigation following the 737 MAX accidents (2018-2019), not in anticipation. In both cases, public doctrine codified, after the fact, the compromises economic and jurisdictional actors had had to impose over the cycle. No structural reason gives agentic healthcare a different trajectory.

These three forces do not coordinate. They converge through network effect. And their convergence defines effective regulation: the set of operational constraints to which an agentic system in healthcare is *actually* subjected, well before doctrine has codified the principles. Economics is not the context of regulation. It is its engine.

10. Conceptual limit cases: PREDICARE and TweenMe

The present text does not argue that these two instances validate the thesis as general proofs. They serve here as conceptual limit cases: testing grounds where causal non-localizability becomes observable and where the regimes described above become operative.

- *PREDICARE*, a sequential predictive medicine architecture (multi-source ingestion, contextual inference, graduated recommendation, longitudinal follow-up), admits a SaMD qualification component by component, but global behavior depends on the trajectory.
- *TweenMe*, a chaining of heterogeneous models (contextual LLM, tabular risk models, Fine & Gray-type survival models, toxicological oracles) with dynamic selection and long memory, produces an emergent value not localizable in any sub-module.

On both terrains, causal non-localizability is observable, the functional stratification regime is insufficient, and the economic arbitrages described above are already at stake. This suffices to invalidate the contrary thesis, according to which the current framework captures as object what it is meant to govern.

11. The auditable orchestration substrate as operational primitive

The proposal of an *auditable orchestration substrate*, defined as a condition of governability, requires being formulated at a level of precision that strictly distinguishes it from existing frameworks (NIST AI RMF, ENISA AI Threat Landscape, ISO/IEC 42001). Failing that, it resembles rebadged common sense, which is what it must not be. The present section traces the discriminating threshold.

The substrate is defined by a single primitive: **behavioral falsifiability**.

Statement. Let T be an execution trajectory of a persistent agentic chain, producing an observable behavior β. The substrate requires that there exist a causal partition {E_k} of T such that:

- (NC, necessary condition) each transition E_k is associated with an identified regulatory role preserved at execution;
- (SC, sufficient condition) the composition of transitions {E_k} allows β to be reconstructed while preserving pertinent causal dependencies, and each transition is inspectable at execution without interrupting the system.

Put differently, the substrate requires that a behavior produced by an agentic chain be reconstructible as an intelligible sequence of attributable transformations. Not merely after the fact, by human interpretation, but during execution itself. Every significant step must leave a causally exploitable trace: which component transformed which information, under which rule, with which responsibility, and in which execution context.

The requirement is close to that of an aeronautical flight recorder, but applied to software decision-making: when a behavior produces a clinical, legal, or prudential effect, it must be possible to retrace the effective trajectory that generated it without depending on an ex post narrative reconstruction.

A chain can therefore be technically performant, compliant with its quality processes, and even regulatorily audited, while remaining behaviorally non-falsifiable if, at the moment it acts, critical context transformations remain neither typed, nor attributable, nor reconstructible.

- *Observable failure criterion.* The substrate is in default if there exists at least one observed behavior β for which no partition {E_k} satisfying (NC) and (SC) can be reconstructed at execution. Doctronic is its direct illustration: the manipulated SOAP note, read by the clinician, does not allow one to trace back to the transition that transformed user content into regulated clinical context. The genealogy has been erased, not by technical defect, but by absence of regulatory typing of the transitions.

- *Operational difference with classical audit.* An ISO/IEC 42001 audit verifies that an organization respects governance processes; it operates on static artifacts (documents, configurations, ex post logs). A substrate verifies that an execution trajectory preserves imputability; it operates on the ongoing dynamic. A system can pass an ISO/IEC 42001 audit and fail the substrate. This is precisely what Doctronic demonstrates: an organization can be compliant with governance standards and deploy a chain whose emergent behaviors are not attributable at execution.

The substrate is not the solution to the problem. It is its necessary condition of solvability: that without which no regulatory doctrine grounded in imputability can apply to a persistent agentic chain. This precision protects the primitive against the confusion between behavioral falsifiability and organizational compliance, which are distinct objects.

12. Articulation with prior work

The present analysis extends three lines: the critique of the LLM-centric paradigm in favor of a composite architecture, the formalization of Clinically-Informed Neural Networks as a class of models internalizing constraints by construction, and the RAISE framework as a doctrine of architectural responsibility in regulated environments.

RAISE assumed that responsibility can only be assigned where it is technically portable. The present thesis completes this assumption with its counterpart: in agentic environments, responsibility can only be assigned where the trajectory is falsifiable. The substrate is the extension of RAISE to an object, the orchestration chain, that RAISE did not thematize explicitly.

13. Conditions of validity of the thesis

Three conditions, and only three, under which the thesis defended here would cease to hold.

- *Firstly*, if a major regulatory overhaul (explicit extension of SaMD to the execution trajectory, or doctrinal formalization of dynamic composition under the AI Act) were to intervene within a short horizon, the prediction of dominance of the third regime would be invalidated. This overhaul is not supported by any current regulatory signal; it remains theoretically possible.
- *Secondly*, if a public technical normalization (for example a CEN-CENELEC or ISO standard on orchestration auditability adopted at European or international scale, legally opposable through harmonization) were to intervene before jurisprudential stabilization, it would render industrial self-regulation secondary and no longer

structuring. This eventuality is compatible with current dynamics around AI auditability standards, but the foreseeable calendar places it downstream, not upstream, of expected jurisdictional decisions.

- *Thirdly*, if causal non-localizability proved empirically circumventable by generalizable dynamic audit mechanisms (for example a standard of regulatory typing of transitions adopted by the industry), the substrate would become observable as a default property, and cumulative divergence would cease to grow. This eventuality is precisely what the present text seeks to provoke by formalizing the substrate as a primitive.

14. Conclusion: structural instability

SaMD remains the central legal anchor for AI devices in healthcare. The AI Act, through its Articles 14, 15, 25, 56, and 95, extends this anchor through solid indirect capture. But neither formalizes the generating trajectory as an autonomous regulatory object, that is, the property that characterizes the class of systems now deployed.

This inadequacy is not a state. It is a trajectory:

Systems are governed where they do not decide, and decide where they are little governed. The gap widens with every cycle of innovation. This is not a divergence: it is a structural instability.

As long as this instability persists, effective regulation will not be produced by doctrine. It will be produced by the insurers-courts-industrials triangle: the three forces that have only ever caught up with doctrine when doctrine missed the real. Three predictions, structural rather than dated, follow.

The regime of industrial self-regulation will become dominant on agentic orchestration in healthcare before any formal doctrinal evolution, codified by the AI Act Codes of Practice and by platform terms of service. Public regulatory doctrine will join it through implementation, not through legislation.

Jurisprudence on PLD 2024/2853 will produce, in the years to come, a few structuring decisions on the imputability of emergent behaviors. These decisions will define the effective civil liability of deployers before the AI Act is revised to codify their principles.

SaMD will survive, but its doctrinal centrality will be relativized. The coming regulatory landscape is one of coexistence: SaMD on devices, AI Act on models, ex post codes on chains. None of these regimes will capture the totality; their articulation will be the regulatory work of the decade now opening.

Agentic orchestration is not, in April 2026, a formal regulatory object. It is already the principal vector of behavior in the architectures it structures, and it will be regulated

(partially, indirectly, and outside doctrine) before doctrine has recognized that it was. This is structural instability.

Public doctrine will not regulate the agentic first. It will codify, later, the compromises imposed by those who had to insure, contract, and judge its accidents.