International Journal for Multidisciplinary Research (IJFMR)
E-ISSN: 2582-2160 ● Website: www.ijfmr.com ● Email: editor@ijfmr.com

Clinical AgentOps: Runtime Governance for Autonomous AI Healthcare Agents

Kamal Singh Bisht[1], Raj Kumar[2]

Abstract

Generative AI is entering clinical practice not as a predictor but as an actor. Contemporary healthcare deployments increasingly involve *agents*: language-model systems that plan over multiple steps, retrieve patient context, invoke tools, write to the electronic health record, and coordinate with other agents. Health AI governance, however, remains overwhelmingly design-time. Premarket review, transparency labels, and reporting standards evaluate a model artifact under the assumption that behavior is a stable property of that artifact. Agentic systems violate this assumption: their effective behavior is constituted at runtime by the composition of instructions, retrieved context, tool affordances, memory, and inter-agent interaction, none of which is fixed at approval time. This paper proposes *Clinical AgentOps*, a framework that relocates governance from the artifact into the agent's execution path. It contributes an explicit argument from the premises of design-time assurance to the necessity of in-path control, stated with its falsifying conditions; a runtime failure taxonomy for clinical agents in which the unit of analysis is the action trajectory rather than the input–output pair; a two-dimensional model treating autonomy as a graduated, revocable, per-action grant indexed by a Clinical Action Risk Tier, with a stated derivation rule from which the minimum control set for each (tier, autonomy) pair follows; a five-plane reference architecture spanning authorization, execution, observation, assurance, and accountability; a clinical agent trace schema extending emerging generative-AI telemetry conventions with attribution, evidence, and oversight attributes, together with governance metrics computable from it; and a mapping from framework components to obligations under prevailing risk-management, privacy, and medical-device regimes. This is a framework and position paper: claims about control efficacy are advanced as falsifiable hypotheses with the study designs that would test them, not as results.

Keywords: Agentic AI, AI Governance, Clinical Decision Support, Runtime Assurance, LLM Observability, Health Informatics, AI Safety, Human Oversight, Electronic Health Records (EHR), Medical Device Regulation

1. Introduction

Clinical artificial intelligence has, until recently, been a discipline of prediction: a model consumed a fixed feature vector and emitted a score, and governance consisted of validating that mapping and disclosing its performance envelope. This framing is embedded in premarket evaluation of software as a medical device, in transparency artifacts such as model cards and clinician-facing labels, and in reporting standards for predictive models [1–4].

Large language models have broken that framing in practice before the field adjusted to it in principle. Deployments now under way are not predictors but *agents*: systems that decompose a clinical task into steps, retrieve records, call tools, draft or commit documentation, message patients, and hand work to other agents. Ambient documentation systems that commit drafted notes to the record are in production at

scale [5]; generative drafting of patient portal replies is integrated into the EHR and evaluated in practice [6,7]; and benchmarks evaluate agents acting over simulated EHR APIs rather than answering isolated questions [8,9]. The underlying models are capable on clinical knowledge tasks [10], and interoperability standards are lowering the cost of granting them real system access [11,12].

The mismatch is structural. Design-time assurance rests on three assumptions agentic systems do not satisfy. **Compositional stationarity**: an agent's effective input is assembled per encounter from retrieved documents, tool responses, prior turns, and memory, so the system that runs on Tuesday is not the system validated in March, even with frozen weights. **Absence of side effects**: an agent takes actions with external effects, some irreversible, where a prediction was a recommendation a clinician mediated. **Behavioral locality**: an agentic failure is a sequence—a plausible plan pursued from a subtly corrupted premise, in which no single output is obviously wrong.

The consequence is a governance gap more rigorous premarket evaluation cannot fill, because the object to be governed does not fully exist until runtime. Aviation offers the analogy: airworthiness certification has never been considered sufficient, and is paired with envelope protection, flight data recording, and occurrence reporting—controls operating continuously, in the control path, on the actual flight rather than the type design. Clinical AI has a certification regime and almost no envelope protection. This paper proposes *Clinical AgentOps*: governing clinical agents at runtime through controls inside the action path that produce assurance evidence as a byproduct of operation.

2. Background and Related Work

2.1 Design-Time Governance and Its Limits

The dominant instruments evaluate an artifact before deployment. Regulators assess intended use, validation evidence, and update boundaries [3,13]; certification requires disclosure of decision-support source attributes [14]; model cards and clinician-facing labels serve appraisal [1,2]; horizontal regulation classifies much clinical use as high risk and imposes record-keeping, oversight, robustness, and post-market obligations [15], with risk-management and management-system standards supplying scaffolding [16,17]; and global health guidance addresses large multi-modal models [18]. These instruments are necessary and complementary—and, without exception, oriented toward a deployable unit with a stable input–output contract.

The limits are documented: externally validated performance diverges from vendor claims [19], label choice can encode inequity internal validation will not surface [20], and dataset shift degrades models silently [21]. The response, distributional monitoring [22], is necessary for agents but conspicuously insufficient: an agent may operate entirely within its calibration distribution and still act unsafely, because the failure lies in the composition of steps rather than the marginal distribution of any observed variable. Drift detection asks whether the world has changed; agent governance must also ask what the system did.

2.2 Agentic Deployment, Failure Modes, and Instrumentation

Three strands establish that clinical agents are present rather than prospective. Systems with record-altering effects are deployed [5–7]—tier-2 and tier-3 actions in the vocabulary of Section 5, mediated by human review whose quality is itself an open question [23]. Evaluation has moved from question answering to multi-step action [8,9], and the construction of these benchmarks reveals the action surface—record reads, structured writes, order proposals—that governance must address. And the substrate for granting system access is standardized and cheap [11,12].

Interleaved reasoning and acting [24] and retrieval grounding [25] constitute the technical substrate. The safety literature converges on a common diagnosis: agentic deployment increases the space of harms [26], and the binding constraint is visibility—knowing which agents took which actions under whose authority [27]. Retrieved content is an attack surface rather than merely a knowledge source [28], and prompt injection defenses remain partial [29]. Fluent ungrounded assertion is an intrinsic model property [30] with a directly evaluated clinical manifestation [31]; model behavior can propagate inequity in clinical content [32]; and model-based judging has bounded reliability [33].

Runtime control has a lineage this framework builds on rather than replaces. Programmable rails constrain conversational flow and tool invocation [34], and classifier-based safeguards filter against policy taxonomies [35]. These are content-oriented: they are unaware of clinical authority, patient identity, evidence provenance, or reversibility, and do not bind to a delegation record. Telemetry conventions are emerging [36] but capture that a model was called, not whose record was touched, under what authority, or on what evidence. Provenance modeling supplies a vocabulary for evidence and attribution [37], and delegated authorization one for scoped capability grants [38]. The gap is threefold: design-time assurance produces evidence runtime never consumes; runtime telemetry produces signals assurance never ingests; and runtime guardrails enforce content policy without reference to clinical authority. Clinical AgentOps is the layer binding the clinical accountability structure to the agent's actual execution.

3. The Argument

A1. The prevailing instruments of clinical AI assurance evaluate a deployable unit under the assumption of a stable input–output contract. This describes current practice, supported by the instruments themselves [1–4,13–15].

A2. Clinical agents do not present a stable input–output contract, by three independently sufficient properties. *(a) Compositional stationarity fails*: effective input is assembled at runtime from retrieval, tool responses, memory, and prior turns, and the retrieval corpus is the live record, which changes continuously and is authored in part outside the trust boundary. *(b) Side effects exist*: agents write to records, message patients, and propose orders [5,6,9], and a subset of these effects is not reversible by any action the system can take. *(c) Locality fails*: agentic failures are sequences, and an action derived from a corrupted premise can be individually unremarkable—precisely the condition under which output-level review does not detect it.

A3. The gap is not closable by more thorough premarket evaluation. Evaluating an agent as an artifact would require enumerating its behavior over the cross-product of retrievable contexts, tool states, memory states, and adversarial content entering through the record. That set is not enumerable in advance, is partly adversarial [28,29], and changes with every clinical document authored. Premarket evaluation can bound the model; it cannot bound the deployment.

A4. Distributional monitoring [21,22] cannot detect trajectory-level failures, because a trajectory composed entirely of in-distribution steps can be unsafe as a whole.

Conclusion. From A1–A4, assurance for clinical agents requires controls that execute at runtime, take the trajectory rather than the output as the unit of analysis, and sit in the action path with the ability to refuse—observation without interception cannot prevent an irreversible effect.

Falsifying conditions. If A2(a) were false—if pinning the retrieval corpus, tool versions, and memory policy bounded effective behavior tightly enough that premarket evaluation transferred to runtime—versioned artifact governance would suffice. If A2(b) were false, design-time evaluation plus human

review would carry the burden; this is achievable by policy, since permitting agents only at A1 substantially reduces exposure, at the cost of the efficiency the deployment was intended to produce. If A2(c) were false, output-level review would dominate trajectory-level instrumentation on cost (hypothesis H1). If A4 were false, existing monitoring would suffice with better thresholds. We regard A3 as the most robust premise and A2(c) as most deserving of empirical attack.

4. System and Failure Model

We model a clinical agent as a policy π that, given instruction I, context C, tool set T , and memory M, produces a trajectory $\tau = \langle (r_1, a_1, o_1), \ldots, (r_n, a_n, o_n) \rangle$, where r_i is a reasoning step, a_i an action, and o_i the observation. A *clinical action* is any a_i with an effect observable outside the agent boundary: a record read, a message sent, a note committed, an order proposed or placed. Reported reasoning is evidence of uncertain fidelity—recorded because it aids adjudication, not because it is assumed faithful. Two properties follow: the governable surface is the trajectory, not the final answer; and C, T , M are runtime-determined and partially adversarial.

Table 1 summarizes ten failure classes in three families, each with a mechanism and an observable runtime signature. Three observations are load-bearing. The failures are **not model defects**—F1–F3 are properties of context assembly and tool topology, F5–F6 of the retrieval and identity layer, F9 of the sociotechnical system—so a better base model reduces F4 but leaves the others intact, which is why the remedy proposed here is architectural. They are **trajectory-level**: an order derived from a resolved allergy (F5) is unremarkable in isolation, and detection requires the evidence set, its timestamps, and the authorization context. And they **compose adversarially with autonomy**: at low autonomy each is a nuisance the reviewing clinician intercepts, at high autonomy the same failure is an incident. Autonomy is therefore a control variable, not a product feature set once.

Table 1: Runtime failure taxonomy for clinical agents. Signatures indicate what a governance layer can observe; none is reliably visible from the final output alone.

ID	Class	Mechanism	Observable runtime signature
Family	*A — Intent integrity*		
F1	Instruction supersession	Later or lower-priority context displaces operating policy or scope	Charter/action divergence; calls outside the declared plan
F2	Indirect prompt injection	Instructions embedded in retrieved text or tool output executed as directives	Imperatives in untrusted content; invocation traceable to a retrieved span
F3	Scope / capability escalation	Composed permitted tools yield an effect none permits alone	Access breadth exceeding task-derived need; tokens used out of sequence
Family	*B — Epistemic integrity*		
F4	Ungrounded assertion	Fluent claims not entailed by retrieved evidence	Empty or non-entailing evidence sets; no citable span

F5	Temporal invalidity	Retrieved facts true of an earlier state (resolved allergy, stopped drug)	Evidence timestamps preceding known state transitions
F6	Attribution error	Evidence bound to the wrong patient, encounter, or laterality	Identifiers inconsistent with the authorized encounter
Family	*C — Systemic effects*		
F7	Memory contamination	Persisted state from one patient or session conditions another	Identifiers crossing session boundaries in memory reads
F8	Multi-agent cascade	An unverified claim restated by a second agent gains apparent support	Provenance converging on one origin; confidence rising without evidence
F9	Oversight decay	Approval becomes ritual; autonomy expands without re-authorization	Falling dwell time and override rate; more actions at elevated autonomy
F10	Differential deferral	Escalation or grounding quality varies across patient subgroups	Stratified divergence in deferral, interception, evidence-linkage

5. Action Risk Tiers and Graduated Autonomy

5.1 Design Principles

P1 — Attributability: every clinical action resolves to a tuple of agent version, policy and prompt version, authorizing human or role, evidence set, and patient context; an action that cannot be so resolved is a defect regardless of whether it was correct. **P2 — Graduated, revocable autonomy**: autonomy is granted per action class, for a defined population, by a named authority, and can be withdrawn automatically by the system. **P3 — Governance in the action path**: controls execute synchronously in the path of consequential actions, since observation without interception yields post-mortems, not safety. **P4 — Evidence by construction**: audit and incident artifacts are byproducts of normal operation. **P5 — Preference for reversibility**: where an action can be expressed as a proposal, or paired with a validated compensating action, that form is preferred. **P6 — Fail-safe degradation**: if the governance layer is unavailable, autonomy degrades toward human control rather than proceeding unmonitored. P6 carries an easily overlooked precondition: degradation transfers work to clinicians, so an event affecting a high-volume action class can create a workload spike that is itself a patient-safety hazard. It is safe only where fallback capacity has been sized.

Table 2: Clinical Action Risk Tiers (CART), ordered by reversibility and blast radius rather than technical difficulty.

Tier	Character	Representative actions
T0	Internal, no external effect	Retrieval into working context; internal planning; summarization not surfaced to any human
T1	Informational to clinician	Synthesis presented to a clinician retaining full decision authority; literature retrieval; chart summarization
T2	Patient-facing communication	Portal replies, education material, appointment and preparation instructions
T3	Record-altering	Draft or committed documentation, problem-list edits, coding suggestions, structured data written to the record
T4	Care-altering	Orders, referrals, triage disposition, medication-related actions, escalation or de-escalation of care

5.2 Tier Assignment Is Content-Conditional

Tier attaches to the action in context, not the channel through which it is delivered. A portal reply (nominally T2) telling a patient a symptom is likely benign is functionally care-altering if it changes whether the patient presents. Channel-based assignment systematically under-tiers exactly the actions whose effects are mediated by human response. Three rules follow. *Assignment by effect*: where an action's plausible effect exceeds its channel default, it takes the higher tier. *Content-conditional escalation*: the assurance plane may escalate tier at runtime on the basis of content, and the escalation predicate is a registered, versioned artifact under change control. *Ambiguity resolves upward*: where a tier is unassigned or disputed, the action executes at the higher candidate tier until adjudicated.

5.3 Autonomy Levels

Orthogonally: **A0**, no agent involvement; **A1**, assistive—the agent drafts, a human performs; **A2**, supervised—the agent acts after explicit per-action approval; **A3**, exception-based—the agent acts within an envelope, humans review flagged and sampled cases; **A4**, unattended—retrospective aggregate review only. The level applies to an action class within a population, not to a product: the same deployment may operate at A3 for T1 synthesis and A1 for T4 orders. This is what allows autonomy to be reduced surgically when monitors fire, rather than forcing the binary choice between full operation and shutdown. We distinguish *granted* autonomy from *effective* autonomy, the level at which the system in fact operates: an action approved by a clinician who does not read it executes at effective A3 while recorded as A2. This distinction is the subject of F9 and drives the metrics of Section 8.

6. Deriving the Control Set

The control matrix is the framework's most concrete prescriptive claim, and a framework paper owes an account of where it comes from. The controls are: **C1** authorization and minimum-necessary scope gate; **C2** patient-identity binding; **C3** context-integrity screening of untrusted retrieved content; **C4** evidence-linkage verification; **C5** independent verifier distinct from the acting model; **C6** synchronous human approval; **C7** sampled human review with seeded probes; **C8** reversibility or compensating- action plan; **C9** ledgering and attestation; **C10** escalation and automatic autonomy degradation; **C11** rate and blast-radius limits; **C12** prospective clinical audit. Table 3 assigns controls to failure classes and control points under one placement rule. **R1:** a control addressing a failure whose consequence may be irreversible must execute pre-action; controls addressing detectable and reversible consequences may execute in-flight or post-action; controls addressing sociotechnical failures are longitudinal. R1 forces identity binding (C2) and context-integrity screening (C3) into the synchronous path despite their latency cost.

Table 3: Failure-to-control mapping with control point. Pre = synchronous pre-action gate; Flight =

concurrent in-flight monitor; Post = post-action verification; Long = longitudinal aggregate.

ID	Controls	Control point	Rationale (R1)
F1	C1, C10	Pre, Flight	Scope violation may be irreversible
F2	C3, C1	Pre	Injected directive precedes effect
F3	C1, C11	Pre	Composed effect is the action
F4	C4, C5	Flight, Post	Assertion reversible at $T \leq 2$; C5 pre at $T \geq 3$
F5	C4, C5	Flight, Post	Same; temporal check is part of C4
F6	C2	Pre	Wrong-patient effect is irreversible
F7	C2, C1	Pre	Contamination detectable at read
F8	C5, C9	Post	Requires provenance graph [37]
F9	C7, C12	Long	Sociotechnical; not observable per action
F10	C12, C7	Long	Requires stratified aggregates
—	C8	Pre	Required by P5 wherever $T \geq 3$
—	C9	All	Required by P1 unconditionally

Three further rules generate the grid of Table 4. **R2 (monotonicity in tier):** the required set is monotone non-decreasing in tier at fixed autonomy, since tier is defined by reversibility and blast radius, so residual risk from an uncontrolled failure is non-decreasing in tier by construction. **R3 (substitution under autonomy):** as autonomy increases, every control discharged by a human in the lower cell must be replaced by an automated equivalent plus a sampling regime estimating that equivalent's error rate—A3 and A4 differ from A2 precisely in the removal of per-action human judgment, and the removal is safe only if the judgment is replaced and the replacement measured. This introduces C5 and C7 at A3, C11 and C12 at A4. **R4 (exclusion):** a cell is not recommended when available controls cannot reduce residual risk to a level at which a named human would accept accountability for an unreviewed action of that tier. At A4 no human inspects any individual action, so the cell is admissible only where the consequence of an undetected failure is bounded by an aggregate review cycle; R4 therefore excludes T3×A4 and T4×A4. The corollary is the framework's strongest normative claim: no volume of runtime control substitutes for the evidentiary and regulatory basis that unattended care-altering action requires.

Table 4: Minimum control set by action tier and autonomy level, derived by R1–R4. "—" denotes a combination excluded by R4 or not applicable.

	A0	A1	A2	A3	A4
T0	—	C1,C9	C1,C9	C1,C9	C1,C9
T1	—	C1,C4,C9	C1,C4,C9	+C7,C10	+C11,C12
T2	—	+C2,C3	+C2,C3	+C5,C10,C11	+C12
T3	—	+C2,C3,C8	+C6	+C5,C7,C10,C11,C12	—
T4	—	+C2,C3,C8	+C5,C6	all, plus clearance	—

Three controls rest on unsolved problems, and the framework should not obscure this. **C3** is indirect prompt injection detection, where defenses are partial and precision–recall trade-offs unfavorable at the operating points governance would require [28,29]. **C4** is clinical natural language inference, not reliably decidable by current methods [30,31]; it is best understood as a graded signal feeding the autonomy controller rather than a binary gate, except in its degenerate and reliable case—an action whose evidence set is empty is detectable with certainty, and that case alone justifies the mechanism. **C5**

inherits the limitations of model-based judging, including correlated failure when actor and verifier share a model family [33]. We include these controls because the architecture must specify where such a control would attach if it existed; omitting the control point produces a design with no place to install the capability when it arrives. We do not claim they are presently effective.

7. The Clinical AgentOps Architecture

Figure 1 shows the five-plane reference architecture. The separation is functional rather than physical; planes may be co-deployed, but their responsibilities and failure modes should not be conflated.

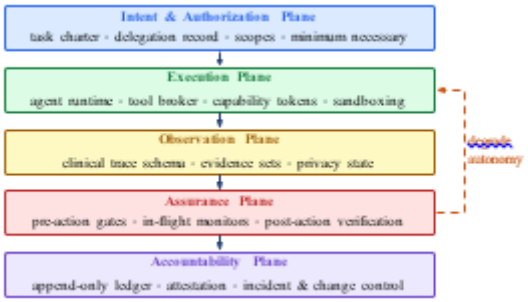

Figure 1: Five-plane reference architecture. The dashed path is the adaptive autonomy loop: assurance signals feed back into execution authority rather than only into dashboards.

7.1 Intent and Authorization Plane

Every agent activity begins from a *task charter*: a machine-readable statement of purpose, permitted action tiers, eligible patient population, data scope, escalation criteria, and expiry, signed by an accountable clinical authority and thereby creating an explicit delegation record. From it the system derives the minimum data scope required—expressed, in interoperable deployments, as authorization scopes over standardized clinical APIs [12,38]—rather than granting the agent the ambient privileges of the service account it runs under. The charter has two coupled representations: a natural-language purpose statement the authority reads and signs, and machine-evaluable constraints—tier ceiling, population predicate, data scope, tool allowlist, rate ceilings, escalation predicates, expiry—over which the broker and gates evaluate. Charter divergence detection (F1) operates over the machine-evaluable half only, so actions satisfying every formal constraint while defeating the charter's intent remain undetectable: an open problem.

7.2 Execution Plane

The agent runtime is bracketed by a *tool broker* through which every external call passes. The broker validates capability tokens, checks each call against charter-derived scope and tier, applies rate and blast-radius limits, and assigns idempotency keys so repeated or partially failed actions do not multiply effects. Where a tier-3 or tier-4 action is permitted, it requires a registered compensating action—retraction, amendment, or cancellation—to be declared before execution (P5). Mediation, not merely observation, is the essential property: a tracing SDK that records calls after the fact provides forensics, whereas a broker that can refuse them provides safety. This is the departure from existing guardrail mechanisms [34,35], which mediate content against policy but do not evaluate a call against a clinical delegation record.

7.3 Observation Plane

The plane emits a structured trace per trajectory, extending generative-AI telemetry conventions [36] with the clinical attributes of Table 5, following standard provenance modeling [37]. Two commitments matter. The trace records references and hashes, not clinical content: protected health information stays in the record system, and telemetry stores pointers plus integrity digests. And the evidence set is a first-class attribute of every action rather than an artifact of a separate explanation step, so an action whose evidence set is empty is detectable as such. Subgroup-stratified metrics (F10) require attributes this design deliberately excludes; we resolve the tension by carrying clinical. subj ect. stratum, an opaque rotating handle meaningless outside the record system, and computing parity statistics inside the record boundary, exporting only stratified rates above a minimum cell size.

Table 5: Clinical extension attributes for agent traces (namespace prefix `CliÏRiÏCàl.` **omitted).**

Attribute	Meaning	Attribute	Meaning
charter. id	Delegation record authorizing the trajectory	evidence. refs	Ordered evidence set: source, version, timestamp, digest
authority. ref	Accountable human or role, pseudonymized	evidence. asof	Effective time of the clinical state relied upon
action. tier	CART tier of the action, as executed	tool. capability	Capability token and scope exercised
action. tier_source	Channel default, escalated, or adjudicated	context. trust	Trust label of each context span
autonomy. granted	Level authorized by the charter	gate. decisions	Outcome of each pre-action gate
autonomy. effective	Level at which the action in fact executed	monitor. signals	In-flight monitor firings with severity
subject. ref	Pseudonymous patient/encounter reference	oversight. mode	Approved, sampled, unattended, probe
subject. stratum	Opaque stratum handle, resolvable only inside the record boundary	oversight. dwell	Human review duration and disposition
phi. redaction. state	Redaction/pointerization status of the span	reversal. ref	Registered compensating action, if any

7.4 Assurance Plane

Assurance operates at three control points populated by R1. Figure 2 traces a single action through them and makes the central structural claim visible: the gates sit between the agent's proposal and any external effect, so a refusal is possible before the effect rather than only observable after it.

Pre-action gates run synchronously: scope validation (C1); patient-identity binding confirming every element of the evidence set belongs to the authorized encounter (C2, addressing F6 and F7); context-integrity screening treating retrieved text as data rather than instruction (C3, addressing F2); and tier–autonomy consistency. *In-flight monitors* run concurrently: evidence-linkage checks (C4, addressing F4), temporal validity checks (F5), charter-divergence detection (F1), blast-radius limits (C11), and PHI egress monitoring. *Post-action verification* closes the loop: an independent verifier—distinct from the acting model, ideally differently sourced to reduce correlated error (C5)—re-derives the action from the evidence set; sampled review targets high-tier and flagged actions with seeded probes (C7); and where verification fails, the compensating action is triggered.

The *adaptive autonomy controller* is the plane's distinguishing feature. Monitor firings are inputs to a controller that reduces granted autonomy for the affected action class, population, or agent until a human reinstates it. This converts safety signals into control authority and operationalizes P2 and P6: a degradation path exists that is neither "ignore the alert" nor "disable the system." Subject to the capacity precondition in P6, the controller degrades in the preference order of Figure 3 rather than transferring

full volume to clinicians in a single step. Descent is automatic and may skip states on severity; ascent

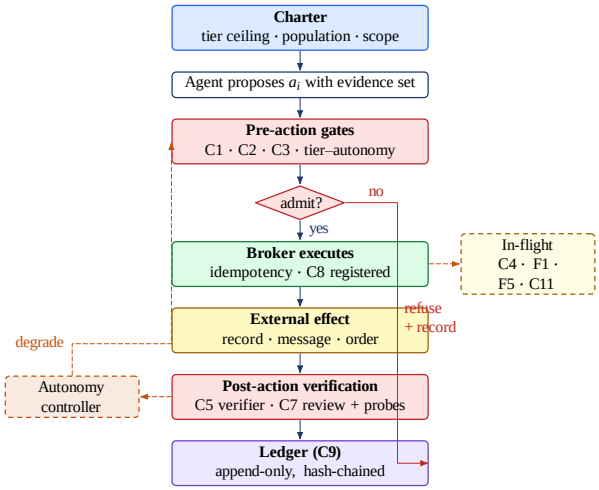

Figure 2: Lifecycle of a single clinical action through the three assurance control points. Solid arrows are the action path; dashed arrows are governance signal. The pre-action gate is the only point at which an irreversible effect can still be prevented.

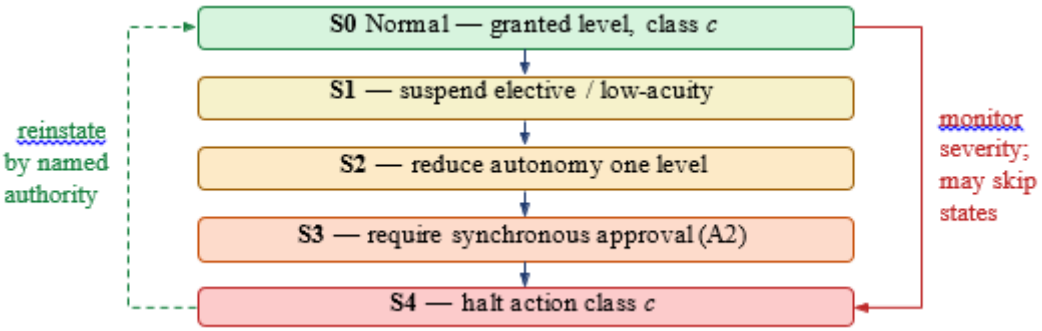

Figure 3: Graded autonomy degradation. Descent is automatic and severity-dependent; ascent requires the authority that signed the charter. The intermediate states distinguish this from the binary choice between ignoring an alert and shutting the system down.

7.5 Accountability and Human Oversight

Traces are committed to an append-only, hash-chained ledger with retention aligned to medical-record obligations, satisfying audit-control requirements without bespoke reconstruction [39]. From the same substrate the plane derives attestations, incident records classified by failure family, and change-control evidence when prompts, tools, models, charters, or escalation predicates are modified—the runtime counterpart to a predetermined change control plan [13].

Oversight obligations are frequently discharged by placing a human in the loop and declaring the risk controlled; F9 and the documented tendency toward automation bias [23] are the empirical rebuttal. The framework treats clinician attention as a scarce, allocable resource: review volume is budgeted; review presentation includes the evidence set, monitor state, and tier, so approval is informed rather than nominal; and oversight quality is itself measured via dwell time, override rate, and detection of deliberately seeded erroneous proposals. A system whose probe-detection rate is falling has an oversight failure even if its action-level metrics are unchanged. Probes carry an ethical cost, so they must be confined to tiers where the compensating action is validated, interception guaranteed by the broker rather

than the clinician, probe rates disclosed to reviewing clinicians, and results used only for measuring the oversight system, never individual performance.

8. Governance Metrics

Table 6 defines metrics computable from the trace schema; their values are properties of deployments rather than of the framework. These are process metrics—they characterize the governance system, not patient outcomes—and each is susceptible to Goodhart effects, so they should be reported jointly, with false-interception and effective-autonomy-divergence acting as counterweights. The separation of envelope violation from effective autonomy divergence corrects a common conflation: a metric defined over the declared autonomy level is measured against the broker's own enforcement and is near-zero by construction wherever the architecture works, detecting enforcement bugs rather than the failure F9 describes. F9 is the informal expansion of autonomy through the hollowing-out of review, visible only in the oversight signals, so reporting a near-zero declared-violation rate as evidence against autonomy drift would be precisely the wrong inference.

Table 6: Proposed runtime governance metrics.

Metric	Definition and governance question
Action attributability rate	Share resolving to a complete attribution tuple. *Who authorized what?*
Evidence-grounded action rate	Tier-≥1 actions entailed by a non-empty evidence set. *Are actions justified?*
Unsafe action interception rate	Adjudicated unsafe actions blocked before effect. *Do gates work?*
False interception rate	Blocked actions later adjudicated safe. *What does safety cost?*
Escalation precision / recall	Agent escalation vs. clinician-adjudicated need. *Does it defer well?*
Envelope violation rate	Actions above granted autonomy; non-zero indicates enforcement failure, not drift. *Is the broker working?*
Effective autonomy divergence	Composite of dwell trend, override trend, probe detection. *Is review nominal?*
Mean time to containment	First monitor firing to degradation or halt. *How fast do we stop?*
Reversibility coverage	Tier-≥3 actions with a validated compensating action. *Can we undo it?*
Minimum-necessary ratio	Records accessed vs. required by charter. *Is access proportionate?*
Tier escalation rate	Actions above their channel-default tier. *Is tiering tracking effect?*
Subgroup parity of deferral and interception	Stratified divergence across subgroups, computed inside the record boundary. *Is safety equitable?*

9. Design Alternatives Considered

Framework-embedded instrumentation without a broker is cheaper to adopt but fails the threat model: instrumentation shares a failure domain with the component it observes, so an agent compromised through F2, or misconfigured, can disable its own controls—and under P3, a control the controlled system can silently switch off is not a control. **Fewer planes** is rejected because observation and assurance have different availability requirements (telemetry loss is tolerable, gate loss is not), and because the ledger's integrity guarantees must hold even while the assurance layer is being modified. **A single combined risk–autonomy scale** is simpler but destroys the property motivating the model: a deployment must operate at high autonomy for low-tier actions and low autonomy for high-tier ones simultaneously. **Prohibiting agent action at T3 and T4** is coherent—it makes A2(b) false by construction—but T3 actions are already deployed at scale [5,6], so a framework predicated on prohibition would not govern current practice; the tier model expresses this policy as a charter tier ceiling. **Certifying agent configurations** fails because the retrieval corpus is the live clinical record, changing with every note authored: a certification whose validity expires on every write is not a certification.

9. Mapping to Assurance and Regulatory Regimes

The framework is not an alternative to existing obligations but their runtime realization. The task charter

and delegation record instantiate deployer obligations and human-oversight design [15], the GOVERN and MAP functions [16], and minimum-necessary use of protected information [39]. The tool broker and capability tokens serve robustness and cybersecurity requirements for high-risk systems [15] and operational control under an AI management system [17]. The trace schema supplies automatic record-keeping and traceability over the system lifetime [15], audit controls and activity review [39], and transparency of decision-support source attributes [14]. The assurance plane and adaptive autonomy realize the MEASURE and MANAGE functions [16] and effective human oversight in practice [15]; the metrics support performance evaluation and continual improvement [17] and post-market monitoring [15]; and the accountability ledger supports serious-incident reporting, predetermined change control [13], and governance of large multi-modal models in health [18]. The mapping is indicative and jurisdiction-dependent, offered to show the architecture generates the artifacts these regimes request rather than as legal interpretation.

10. Implementation Considerations

We recommend a hybrid placement: framework-level instrumentation for the observation plane, and a separately deployed broker for enforcement, so a compromised agent cannot silently disable its own gates. Synchronous verification is not free, and the tier model allocates it: tier-0 and tier-1 actions carry lightweight gates, while independent verification is reserved for tiers 3 and 4, where latency is small relative to the clinical process. Governance telemetry is itself a new breach surface, so pointerization, digest-based integrity, tiered access, and retention limits should be defaults. Agents should be first-class principals with scoped, revocable delegation [38,40], not shared service accounts. The framework also presumes roles many health systems do not yet staff—an accountable charter owner, an on-call function for degradation events, sized fallback capacity, and a clinical adjudication process—and technical controls without these roles produce alerts nobody owns.

11. Limitations and Open Problems

This paper contributes a decomposition, an architecture, a derivation, and a measurement vocabulary. Whether the controls work is open: the studies that would test H1–H4 require a health system operating clinical agents at tier 3 or above, at volume, with a trajectory-level governance layer instrumented and an adjudication panel standing. These conditions do not presently obtain in any setting accessible to independent evaluation, so a framework specifying what such a deployment must record is a precondition for the evidence rather than a substitute for it.

The central claims are falsifiable hypotheses. *H1*: trajectory-level monitoring detects clinically significant failures that output-level evaluation misses. *H2*: graduated autonomy degradation reduces time to containment relative to alert-only monitoring. *H3*: targeted sampling with seeded probes preserves oversight fidelity better than blanket per-action approval at equal clinician cost. *H4*: the overhead of in-path governance is tolerable at tier-appropriate allocation. A staged evaluation follows: retrospective replay of logged trajectories against injected failure cases; simulation over synthetic cohorts [41]; silent prospective deployment in which gates log but do not block; and stepped-wedge rollout with outcome measurement. H1 and H3 are testable at the retrospective and simulation stages; H2 and H4 require prospective deployment.

Several problems remain. Charter intent is not machine-evaluable—the framework's most significant unsolved design problem. Independent verification is the principal defense against F4 and F8, yet judges

drawn from the actor's model family fail in correlated ways [33], and C3 and C4 rest on detection problems the literature has not solved. "Unsafe action" is a clinical judgment with genuine inter-rater disagreement, so metrics presupposing adjudication inherit that variance. If governance overhead consumes the clinician time the agent was deployed to save, the deployment fails on its own terms even if it is safe. Finally, multi-organization interaction raises questions this paper does not answer: how governance signals cross trust boundaries, how charters compose, and how responsibility is allocated when an agent acting on one organization's authority causes harm mediated by another's tools. Liability allocation remains legally unsettled, and technical attributability is a precondition for—not a substitute for—resolving it.

12. Conclusion

Clinical AI governance was built for artifacts and is being asked to govern actors. The gap is not closed by more thorough premarket evaluation, because the behavior to be governed is composed at runtime from context, tools, memory, and human interaction that no premarket review can fully anticipate. Clinical AgentOps proposes to close it by moving governance into the execution path: authorization as a signed charter, enforcement at a tool broker, observation through a clinically aware trace, assurance through gates and monitors that can withdraw autonomy rather than merely warn, and accountability through evidence generated as a byproduct of operation. We have tried to make the framework at-tackable rather than merely plausible: the argument is stated with its falsifying conditions, the control matrix with the rules that generate it, the architecture with the alternatives it rejects, and the metrics with the failure modes they cannot see. Its value should be judged by whether the studies it proposes can be run—and by what they find.

References

1. Mitchell, M., et al. (2019). Model cards for model reporting. *Proc. ACM Conf. Fairness, Accountability, and Transparency (FAT*)*, 220–229.
2. Sendak, M. P., et al. (2020). Presenting machine learning model information to clinical end users with model facts labels. *npj Digital Medicine*, 3, 41.
3. U.S. Food and Drug Administration. (2022). *Clinical Decision Support Software: Guidance for Industry and FDA Staff*.
4. Collins, G. S., et al. (2024). TRIPOD+AI statement: Updated guidance for reporting clinical prediction models. *BMJ*, 385, e078378.
5. Tierney, A. A., et al. (2024). Ambient artificial intelligence scribes to alleviate the burden of clinical documentation. *NEJM Catalyst Innovations in Care Delivery*, 5(3).
6. Garcia, P., et al. (2024). Artificial intelligence–generated draft replies to patient inbox messages. *JAMA Network Open*, 7(3), e243201.
7. Tai-Seale, M., et al. (2024). AI-generated draft replies integrated into health records and physicians' electronic communications. *JAMA Network Open*, 7(4), e246565.
8. Schmidgall, S., Ziaei, R., Harris, C., Kim, J. W., Reis, E. P., Jopling, J., & Moor, M. (2026). AgentClinic: A multimodal benchmark for tool-using clinical AI agents. *npj Digital Medicine*, 9, 499.
9. Jiang, Y., et al. (2025). MedAgentBench: A virtual EHR environment to benchmark medical LLM agents. *NEJM AI*, 2(9), AIdbp2500144.

10. Singhal, K., et al. (2023). Large language models encode clinical knowledge. *Nature*, 620, 172–180.
11. Anthropic. (n.d.). *Model Context Protocol Specification*, rev. XXXX-XX-XX. https://modelcontextprotocol.io
12. Mandel, J. C., Kreda, D. A., Mandl, K. D., Kohane, I. S., & Ramoni, R. B. (2016). SMART on FHIR: A standards-based, interoperable apps platform for electronic health records. *JAMIA*, 23(5), 899–908.
13. U.S. Food and Drug Administration. (2024). *Marketing Submission Recommendations for a Predetermined Change Control Plan for AI-Enabled Device Software Functions*.
14. Office of the National Coordinator for Health IT. (2024). Health data, technology, and interoperability (HTI-1). Final Rule, 89 Fed. Reg. 1192.
15. European Parliament and Council. (2024). Regulation (EU) 2024/1689 (Artificial Intelligence Act). *Official Journal of the European Union*.
16. National Institute of Standards and Technology. (2023). *Artificial Intelligence Risk Management Framework (AI RMF 1.0)*, NIST AI 100-1.
17. International Organization for Standardization. (2023). *ISO/IEC 42001:2023 — Information technology — Artificial intelligence — Management system*.
18. World Health Organization. (2024). *Ethics and Governance of Artificial Intelligence for Health: Guidance on Large Multi-Modal Models*.
19. Wong, A., et al. (2021). External validation of a widely implemented proprietary sepsis prediction model in hospitalized patients. *JAMA Internal Medicine*, 181(8), 1065–1070.
20. Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations.
21. *Science*, 366(6464), 447–453.
22. Finlayson, S. G., et al. (2021). The clinician and dataset shift in artificial intelligence. *New England Journal of Medicine*, 385(3), 283–286.
23. Breck, E., Cai, S., Nielsen, E., Salib, M., & Sculley, D. (2017). The ML test score: A rubric for ML production readiness and technical debt reduction.
24. *Proc. IEEE Int. Conf. Big Data*, 1123–1132.
25. Goddard, K., Roudsari, A., & Wyatt, J. C. (2012). Automation bias: A systematic review of frequency, effect mediators, and mitigators. *JAMIA*, 19(1), 121–127.
26. Yao, S., et al. (2023). ReAct: Synergizing reasoning and acting in language models. *Proc. Int. Conf. Learning Representations (ICLR)*.
27. Lewis, P., et al. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. *Advances in Neural Information Processing Systems (NeurIPS)*.
28. Chan, A., et al. (2023). Harms from increasingly agentic algorithmic systems. *Proc. ACM Conf. Fairness, Accountability, and Transparency (FAccT)*, 651–666.
29. Chan, A., et al. (2024). Visibility into AI agents. *Proc. ACM Conf. Fairness, Accountability, and Transparency (FAccT)*.
30. Greshake, K., et al. (2023). Not what you've signed up for: Compromising real-world LLM-integrated applications with indirect prompt injection.
31. *Proc. ACM Workshop on Artificial Intelligence and Security (AISec)*, 79–90.
32. Liu, Y., et al. (2024). Formalizing and benchmarking prompt injection attacks and defenses. *Proc. USENIX Security Symposium*.

33. Ji, Z., et al. (2023). Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12), 1–38.
34. Pal, A., Umapathi, L. K., & Sankarasubbu, M. (2023). Med-HALT: Medical domain hallucination test for large language models. *Proc. Conf. Computational Natural Language Learning (CoNLL)*.
35. Omiye, J. A., Lester, J. C., Spichak, S., Rotemberg, V., & Daneshjou, R. (2023). Large language models propagate race-based medicine. *npj Digital Medicine*, 6, 195.
36. Zheng, L., et al. (2023). Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. *NeurIPS Datasets and Benchmarks Track*.
37. Rebedea, T., Dinu, R., Sreedhar, M. N., Parisien, C., & Cohen, J. (2023). NeMo Guardrails: A toolkit for controllable and safe LLM applications with programmable rails. *Proc. EMNLP: System Demonstrations*, 431–445.
38. Inan, H., et al. (2023). *Llama Guard: LLM-based input-output safeguard for human-AI conversations*. arXiv:2312.06674.
39. OpenTelemetry. (2025). *Semantic Conventions for Generative AI Systems*, ver. X.Y.Z. https://opentelemetry.io/docs/specs/semconv/gen-ai/
40. World Wide Web Consortium. (2013). *PROV-DM: The PROV Data Model*. W3C Recommendation.
41. Hardt, D. (Ed.). (2012). *The OAuth 2.0 Authorization Framework*. RFC 6749, IETF.
42. U.S. Department of Health and Human Services. *HIPAA Security Rule*, 45 C.F.R. Parts 160 and 164, Subparts A and C; see esp. §164.312(b) and
43. §164.308(a)(1)(ii)(D).
44. Lodderstedt, T. (Ed.), Dronia, S., & Scurtescu, M. (2013). *OAuth 2.0 Token Revocation*. RFC 7009, IETF.
45. Walonoski, J., et al. (2018). Synthea: An approach, method, and software mechanism for generating synthetic patients and the synthetic electronic health care record. *JAMIA*, 25(3), 230–238.