Article

Not peer-reviewed version

From Prediction to Agency: A Constrained Decision Framework and Governance Stack for Agentic AI in Clinical Diagnostics

Yunguo Yu *

Posted Date: 12 May 2026

doi: 10.20944/preprints202605.0723.v1

Keywords: agentic AI; clinical decision support; constrained decision processes; AI governance; agency non-transferability; policy admissibility; compositional risk; autonomy levels; patient safety

 doi:10.20944/preprints202605.0723.v1

Article

From Prediction to Agency: A Constrained Decision Framework and Governance Stack for Agentic AI in Clinical Diagnostics

Yunguo Yu

Zyter|TruCare, 8000 Towers Crescent Dr, Vienna, VA 22182, United States; yuyunguo@gmail.com

Abstract

Clinical artificial intelligence systems are transitioning from predictive tools that generate diagnostic outputs for human interpretation to agentic systems capable of autonomous multi-step action within clinical workflows, including ordering laboratory tests, initiating medication reconciliation, and updating patient records. Existing trust frameworks, designed for advisory systems and built on output verification and confidence calibration, do not address the governance requirements of autonomous action. We identify an agency gap: the structural mismatch between validated predictions and unvalidated action policies. Using a partially observable constrained decision process (PO-CDP) formalism, we establish the principle of agency non-transferability, demonstrating that trust calibrated at the diagnostic level does not imply safe or appropriate action policies under real-world clinical, institutional, and legal constraints. To address this gap, we propose a three-layer governance stack—epistemic soundness, policy safety, and institutional traceability—that provides verifiable guarantees at each stage of the agentic decision pipeline. This paper presents a theoretical governance framework; the phase-specific milestones in the backcasting roadmap define the empirical validation agenda for each deployment stage. A compositional risk analysis formally predicts that individually safe components can produce unsafe system-level behavior through nonlinear error propagation. An extended backcasting roadmap defines three empirically testable phases for the transition to governed agentic systems: sandboxed action proposals (2027–2029), credentialed policy systems (2030–2032), and supervised autonomy (2033–2035). The transition to agentic clinical AI constitutes a paradigm shift from prediction correctness to policy safety under constraint, requiring institutional design rather than technical improvement alone.

Keywords: agentic AI; clinical decision support; constrained decision processes; AI governance; agency non-transferability; policy admissibility; compositional risk; autonomy levels; patient safety

1. Introduction

The integration of artificial intelligence (AI) into clinical medicine has progressed through three identifiable phases. The first generation comprised classification systems that mapped patient data to diagnostic labels, evaluated primarily on accuracy metrics against curated benchmarks [1,2]. The second generation introduced recommendation systems that integrated clinical guidelines and confidence estimates, enabling calibrated outputs with explicit uncertainty quantification [3,4]. A third generation is now emerging: agentic systems capable of multi-step reasoning, tool use, and autonomous interaction with electronic health record (EHR) systems [5,8,10,25]. These systems do not merely suggest diagnoses; they retrieve structured clinical data from institutional processes and business rules [9], propose and, in some configurations, execute sequences of clinical actions—ordering laboratory tests, initiating medication reconciliation, scheduling specialist referrals, and updating patient records.

This transition from prediction to agency has consequences for trust that current frameworks were not designed to address, and the gap between computational capability and clinical evidence

 doi:10.20944/preprints202605.0723.v1

for agentic systems remains substantial [5,25]. A recent scoping review of agentic AI in health care identified 47 studies spanning clinical decision support, workflow automation, and patient monitoring, yet found that none included formal governance mechanisms for autonomous action—highlighting the gap this paper addresses [25]. The dominant trust paradigm in clinical AI rests on output verification: the validation of diagnostic predictions against reference standards, calibrated confidence scores, and guideline concordance [3,11,12]. A recent backcasting analysis identified three temporal pivot points for clinician adoption of AI diagnostics—verifiable AI by 2030, agentic governance by 2035, and futures literacy by 2040—and proposed a dual-process architecture in which a large language model (LLM) generates diagnostic hypotheses that are verified in real time by a locally deployed small language model (SLM), producing a calibrated confidence score [4]. Empirical support for this calibration approach was demonstrated in a validation study of 6,689 cardiovascular cases from the Medical Information Mart for Intensive Care III (MIMIC-III) dataset, in which a dynamic confidence and transparency scoring framework reduced clinician override rates to 33.3%, with high-confidence predictions (90–99% confidence) overridden at only 1.7% [3].

Such calibration results are encouraging for advisory systems but insufficient for agentic ones. The output verification framework carries an implicit assumption that correct predictions will lead to safe clinical decisions. This assumption is structurally flawed in the agentic context. A system may produce a diagnostically correct output—a well-calibrated posterior probability for community-acquired pneumonia—while simultaneously proposing an action sequence that is clinically inappropriate, institutionally unauthorized, or legally impermissible: ordering a bronchoscopy without documented indication, prescribing a contraindicated antibiotic, or initiating a referral pathway outside the institution's scope-of-practice framework. The prediction is sound; the policy is not.

This mismatch constitutes what we term the *agency gap*: the structural divergence between validated predictions and unvalidated action policies. The agency gap is not a deficiency in model performance amenable to additional training data or architectural refinement. It is a property of the transition from prediction to action, arising because the verification requirements for diagnostic outputs (epistemic validity) are categorically different from those for clinical actions (policy admissibility). Diagnostic verification asks whether a hypothesis is correct; action verification asks whether a proposed intervention is *permissible* under the constraints governing clinical practice.

This paper extends the previously published 2040 backcasting roadmap [4] by addressing the governance gap that emerges when advisory AI systems begin to act autonomously—specifically, the underdefined transition between the 2030 verifiable AI pivot and the 2035 agentic governance pivot. We make four contributions. First, we formalize clinical AI agency as a partially observable constrained decision process (PO-CDP), providing a rigorous foundation for analyzing agentic behavior under real-world clinical constraints. Second, we establish a theoretical claim—*agency non-transferability*—demonstrating that trust calibrated at the diagnostic level does not transfer to the action level. Third, we propose a three-layer governance stack that operationalizes distinct verifiable guarantees: epistemic soundness, policy safety, and institutional traceability. Fourth, we extend the backcasting roadmap with empirically testable milestones for the transition to governed agentic systems, and analyze the compositional risks that emerge when verified components interact within an agentic pipeline.

2. Formalizing Clinical AI as a Constrained Decision Process

2.1. From Prediction to Decision Processes

Most clinical AI systems deployed to date are framed as *predictors*: they map patient data (x) to a diagnostic label, risk score, or recommendation. This paradigm implicitly treats clinical decision-making as a single-step inference problem, where correctness can be evaluated against a ground-truth label or guideline concordance.

Real clinical care, however, is inherently sequential, context-dependent, and action-oriented. A diagnosis is not an endpoint but an intermediate belief that informs a sequence of interventions—ordering tests, initiating treatments, consulting specialists, and adjusting plans over time. As AI

 doi:10.20944/preprints202605.0723.v1

systems evolve to participate in or automate these workflows, they must be understood not as predictors but as *decision-making entities operating over time*.

This shift is naturally captured by the formalism of a Markov Decision Process (MDP) [13], which models decision-making as a sequence of states, actions, and outcomes. In this framework, the state (s_t) encodes the patient's clinical condition; the action (a_t) represents a clinical intervention; the transition function ($P(s_{t+1} \mid s_t, a_t)$) captures how the patient state evolves in response to interventions; and the reward function ($R(s_t, a_t)$) encodes clinical objectives such as improved outcomes, reduced risk, and efficient resource utilization.

While the MDP provides a useful abstraction, it is insufficient for clinical contexts because it assumes full observability and unconstrained optimization—assumptions that do not hold in real-world medicine.

2.2. The Clinical PO-CDP: Incorporating Partial Observability and Constraints

Clinical decision-making is partially observable: no clinician or AI system has access to the true underlying patient state. Decisions rely on incomplete, noisy observations from EHRs, patient histories, and diagnostic tests [15]. Clinical actions are not chosen solely to maximize expected reward; they are bounded by evidence-based guidelines, institutional policies such as formularies and care pathways, scope-of-practice regulations, and legal and ethical constraints including liability, consent, and safety standards.

To capture these realities, we extend the classical MDP into a Partially Observable Constrained Decision Process (PO-CDP) [14]:

- Observation $o_t \sim P(o_t \mid s_t)$: incomplete view of patient state
- Belief $b_t = P(s_t \mid o_{1:t})$: probabilistic estimate of the true state
- Action $a_t \in \mathcal{A}$: candidate interventions
- Constraint set $\mathcal{C}$: defines admissible actions under context

The defining feature of the PO-CDP is that *optimality is subordinate to admissibility*. Even if an action has high expected clinical benefit, it may be impermissible if it violates constraints. Formally, the set of valid actions is:

$$a_t \in \mathcal{A}_{\text{valid}}(b_t, \mathcal{C}) \tag{1}$$

where $\mathcal{A}_{\text{valid}} \subseteq \mathcal{A}$ is the subset of actions that satisfy all applicable constraints given the current belief state.

This formulation introduces a key distinction: clinical decision-making is constrained policy selection under uncertainty, not unconstrained optimization. The constraint set $\mathcal{C}$ is multi-dimensional, encompassing clinical constraints (evidence-based guidelines, contraindications, disease-specific pathways), institutional constraints (formulary restrictions, resource availability, local care protocols), scope-of-practice constraints (role-based permissions, credentialing limits), and legal and ethical constraints (liability frameworks, consent requirements, safety regulations).

2.3. Two Distinct Verification Problems

The transition from prediction to constrained decision-making reveals two distinct verification problems.

The first is *output verification* (epistemic validity): given input x, is the predicted hypothesis H correct? This is the domain of diagnostic accuracy, calibration, and guideline concordance. It can be expressed probabilistically as $P(H \mid x, G)$, where G denotes the guideline or knowledge base. This corresponds to the verification layer established in prior work [3,4].

The second is *action verification* (policy admissibility): given a belief about the patient state, is the proposed action or sequence of actions permissible and safe? This requires evaluating constraint satisfaction, contextual appropriateness, and downstream consequences. Unlike output verification, which is primarily statistical, action verification is contextual, relational, and institutional.

 doi:10.20944/preprints202605.0723.v1

2.4. The Agency Gap and Agency Non-Transferability

A central implication of the PO-CDP framework is that trust in predictions does not imply trust in actions. Before formalizing this claim, we distinguish two related concepts. The *agency gap* is the problem: the structural mismatch between validated predictions and unvalidated action policies. *Agency non-transferability* is the principle explaining why the gap exists: the optimization objectives for diagnostic correctness and action admissibility are structurally independent, operating over different information and constraint spaces.

Proposition (Agency Non-Transferability). *There exist PO-CDPs in which a policy maximizing epistemic soundness selects actions that violate constraints in $\mathcal{C}$.*

Proof sketch, by counterexample. Consider a PO-CDP with state space $S = \{s_1, s_2\}$, action space $A = \{a_1, a_2\}$, and a single observation o yielding belief $b(s_1) = 0.95$. The epistemically optimal diagnosis is $H = s_1$ with high confidence. Let the reward function satisfy $R(s_1, a_1) > R(s_1, a_2)$, so unconstrained optimization selects a_1. Now define constraint $c \in \mathcal{C}$: "action a_1 is impermissible when institutional authorization K is absent." Because $\mathcal{C}$ depends on an exogenous factor K (e.g., credentialing status, formulary restriction) not represented in the diagnostic mapping (x, H), the epistemically optimal policy may select a constraint-violating action. Formally, $\arg\max_a R(s_1, a) \notin \mathcal{A}_{\text{valid}}(b, \mathcal{C})$ when K is absent. This establishes that the two optimization problems—output verification and action admissibility—are structurally independent: satisfying one provides no guarantee about the other. □

The practical consequence is that a system may correctly infer a diagnosis (high epistemic validity) yet propose or execute an action that is unnecessary (over-testing), unsafe (contraindicated treatment), unauthorized (outside scope of practice), or inefficient (resource misuse). This occurs because diagnosis operates over beliefs whereas actions operate over policies constrained by context, including institutional, legal, and scope-of-practice factors absent from the diagnostic objective. Trust must therefore be established at two distinct levels: what the system believes (diagnostic correctness) and what the system does (policy safety and admissibility). Failure to distinguish these leads to a systematic underestimation of risk in agentic systems.

2.5. Clinical AI as a Safety-Critical Control System

The PO-CDP framing situates clinical AI within a broader class of safety-critical systems, aligning it with disciplines beyond traditional machine learning. From control theory, we inherit the notion of systems that must remain stable and safe under uncertainty and disturbance. From formal verification, we adopt the requirement that system behaviors satisfy specified constraints under all admissible conditions. From AI alignment research, we recognize that optimizing for a reward function alone can produce unintended or unsafe behaviors, including specification gaming and reward hacking [17,18]. Recent work on constrained policy optimization and alignment techniques—including reinforcement learning from human feedback (RLHF) [28] and constitutional AI [27]—demonstrates that incorporating explicit oversight into the training loop can reduce harmful outputs, providing empirical support for the constrained decision-making paradigm advanced here. Constrained reinforcement learning methods have shown that incorporating structural constraints into policy optimization improves both performance and constraint compliance in operational domains [7].

Within this context, the action validation layer introduced in the following section can be read as a form of *runtime assurance* [19]: a mechanism that enforces constraint satisfaction at execution time, regardless of the underlying model's internal reasoning. Current clinical AI evaluation is model-centric, focused on accuracy, area under the receiver operating characteristic curve (AUROC), and calibration curves [16]. These metrics are appropriate for prediction tasks but insufficient for agentic systems. Evaluation must shift from model performance to policy behavior under constraints: the frequency of constraint violations, appropriateness of action sequences, robustness under distribution shift, and alignment with institutional policy.

 doi:10.20944/preprints202605.0723.v1

3. The Governance Stack as a System of Verifiable Guarantees

3.1. From Conceptual Framework to Enforceable Guarantees

The formalization of clinical AI as a PO-CDP establishes that safe deployment requires more than accurate predictions: it requires guarantees over system behavior. Three properties must hold: (1) epistemic soundness, meaning the system's beliefs about the patient are justified; (2) policy safety, meaning the system's actions are permissible under constraints; and (3) institutional traceability, meaning the system's decisions are auditable and attributable. Existing clinical AI systems predominantly address the first property. As systems evolve toward agentic behavior, the latter two become equally critical. This section introduces a three-layer governance stack that operationalizes these guarantees as verifiable system properties rather than aspirational design goals (Figure 1).

3.2. Layer 1: Epistemic Soundness (Output Verification)

The first layer ensures that the system's diagnostic or inferential outputs are epistemically justified. Formally, this corresponds to estimating:

$$P(H \mid x, G) \tag{2}$$

where H is a clinical hypothesis, x is observed patient data, and G is a structured knowledge base. This layer can be implemented using the dual-process architecture proposed in prior work [4]: a generative model proposes hypotheses, and a verification model evaluates consistency with evidence and guidelines, producing a calibrated confidence estimate. Epistemic soundness can be empirically assessed through calibration curves, error rates stratified by confidence tiers, and guideline concordance metrics.

Crucially, epistemic soundness covers belief formation only, not action selection. A system may produce a highly reliable diagnosis yet propose an unsafe or impermissible intervention. This limitation motivates the second layer. Diagnostic outputs that fall below a pre-specified confidence threshold are routed to mandatory clinician review rather than forwarded to Layer 2 for action proposal, reflecting the principle that epistemic uncertainty is a sufficient condition for human escalation independent of constraint evaluation.

3.3. Layer 2: Policy Safety (Action Validation)

The second layer addresses the central challenge of agentic systems: ensuring that proposed or executed actions are permissible, contextually appropriate, and safe. Given a belief state b_t and a constraint set $\mathcal{C}$, an action a_t is admissible if and only if it satisfies Eq. 1. This layer enforces a policy admissibility condition, transforming the system from an unconstrained optimizer into a constraint-governed decision-maker.

The action validation layer acts as a runtime constraint filter, evaluating each proposed action before execution. The mechanism draws on safety filters from control theory, constraint enforcement in formal verification, and recent guardrail architectures for LLM-based agents [18,29]. Unlike rule-based clinical decision support systems, which operate as static alerts, this layer operates dynamically at decision time, evaluates sequences of actions rather than isolated rules, and integrates patient-specific context and institutional policy.

We propose that the constraint set $\mathcal{C}$ be represented as a hybrid architecture combining two components. The first is a *rule engine* encoding deterministic institutional policies—formulary restrictions, scope-of-practice boundaries, and absolute contraindications—as first-order logic predicates evaluated in constant time per action. The second is a *temporal logic monitor* encoding sequential constraints—e.g., "do not prescribe anticoagulant X within 72 hours of procedure Y"—as linear temporal logic (LTL) formulas evaluated over proposed action sequences. For a proposed action sequence $\langle a_t, a_{t+1}, \ldots, a_{t+k} \rangle$, the admissibility check evaluates each action against the rule engine and the full sequence against the temporal logic monitor, yielding a binary admissibility verdict plus a structured explanation of any

 doi:10.20944/preprints202605.0723.v1

violated constraints. Conflict resolution between overlapping constraints follows a priority ordering: legal constraints override institutional policies, which override clinical guidelines.

Policy safety can be evaluated through constraint violation rate, frequency of overridden or rejected actions, contextual appropriateness scores assessed by expert adjudication, and outcome-aligned appropriateness measures such as unnecessary test rates and adverse events. This layer directly addresses the agency gap identified in Section 2 by ensuring that even if a system's beliefs are correct, its actions must still pass an independent admissibility test.

A constraint qualifies as *hard* and therefore in-scope for the rule engine if it is (1) binary and context-independent—either satisfied or violated with no intermediate states—and (2) derivable from a closed, auditable source such as a formulary list, a credentialing database, or a statutory requirement. Constraints that are graded, probabilistic, or dependent on clinical judgment—for example, risk-stratified dosing recommendations or context-sensitive referral thresholds—are classified as *soft* and are excluded from the rule engine by design. Soft constraints represent clinical judgment territory that the governance stack does not attempt to automate. When a proposed action falls into a region that is neither clearly permitted nor clearly prohibited under the hard constraint set—for instance, a dose that approaches but does not cross an absolute contraindication threshold—the system defaults to a conservative fallback: the action is escalated to human review rather than permitted or blocked autonomously. Probabilistic guideline content that requires weighing competing risk estimates is handled upstream by the Layer 1 calibration module, not the Layer 2 constraint engine, preserving the separation of epistemic and policy verification.

Computational Implementation

The rule engine and temporal logic monitor both have established computational counterparts. The rule engine is implementable as a RETE-based forward-chaining production rule system—for example, Drools or OpenL Tablets—which provides amortized O(1) per-action match propagation for static constraint sets, satisfying the real-time decision support requirement. The LTL monitor for sequential constraint evaluation maps directly onto runtime verification frameworks: SPOT (automata-theoretic LTL model checking), Montre (timed regular expressions over clinical event streams), or equivalent tools that compile LTL formulas into finite automata and evaluate them incrementally over action traces. Constraint conflict resolution following the legal-over-institutional-over-clinical priority ordering is implementable as a lexicographic priority constraint satisfaction problem, solvable in polynomial time for the bounded constraint sets typical of clinical policy. A proof-of-concept implementation of the Layer 2 constraint engine will be made available upon acceptance.

3.4. Layer 3: Institutional Traceability (Accountability Infrastructure)

The third layer ensures that all decisions made by the system are fully traceable, auditable, and attributable within an institutional context. We define a traceability function:

$$T : (b_t, a_t) \rightarrow \mathcal{G}_{\text{prov}} \tag{3}$$

where $\mathcal{G}_{\text{prov}}$ is a provenance graph capturing the full decision context. Each decision record includes input data and derived features, the inferred belief state, proposed actions and alternatives, constraint evaluations, confidence estimates, and execution outcomes.

This layer enables three institutional functions. First, auditability: retrospective review of decisions and comparison against outcomes. Second, accountability: attribution of decisions across actors. Third, performance monitoring: detection of calibration drift and identification of systematic errors.

Agentic systems introduce a shift from individual to *distributed responsibility*. Decisions are no longer attributable to a single clinician or a standalone device, but to a network of actors: the clinician providing oversight and judgment, the institution defining policy and governance, the AI system executing decision support, and the infrastructure providing verification and logging. We term this *distributed clinical agency*: accountability is distributed but must remain explicit and reconstructable.

 doi:10.20944/preprints202605.0723.v1

3.5. Integration Across Layers

The three guarantees operate sequentially and interdependently. Epistemic soundness produces a validated belief; policy safety filters admissible actions; institutional traceability records and evaluates outcomes. Failure at any layer has distinct implications (Table 1).

Table 1. Failure modes across governance layers, with detection methods and mitigation strategies.

Layer	Failure Mode	Consequence	Detection	Mitigation
Epistemic	Incorrect belief	Misdiagnosis	Calibration drift monitoring; stratified error rates	Retrain; revert to backup model
Policy	Invalid action	Unsafe intervention	Constraint violation logs; override frequency audit	Block execution; route to clinician
Traceability	Missing audit trail	Unassignable responsibility	Completeness checks; gap detection algorithms	Expand logging infrastructure

The governance stack differs from prior approaches in several respects. Clinical decision support systems focus on alerts and recommendations but do not enforce action admissibility. Model validation frameworks evaluate predictive performance but do not constrain downstream behavior. AI safety mechanisms typically operate at the model level, whereas this stack operates at the system and institutional level. Together, these layers transform clinical AI from a predictive tool into a governed agentic system.

3.6. Worked Example: Governance Stack in Action

Consider an agentic system managing post-operative anticoagulation after total knee arthroplasty. The system receives patient data x: a 67-year-old female, post-operative day 1, history of atrial fibrillation on apixaban, normal renal function, and a hemoglobin drop from 12.1 to 10.3 g/dL.

Layer 1 (Epistemic Soundness). The system infers belief b_t: high risk of venous thromboembolism (VTE) given recent arthroplasty and atrial fibrillation, with moderate bleeding risk given hemoglobin decline. The diagnostic output—"restart anticoagulation; consider VTE prophylaxis dose"—is epistemically justified with calibrated confidence of 0.89.

Layer 2 (Policy Safety). The system proposes action sequence: (1) restart apixaban 5 mg twice daily, (2) order repeat complete blood count in 24 hours. The rule engine evaluates: apixaban is on formulary (✓); dose is consistent with renal function (✓); prescribing provider has credentialing for anticoagulant management (✓). The temporal logic monitor evaluates: no sequential constraint violated (e.g., no contraindicated procedure within 72 hours). The action sequence passes admissibility.

Now consider a variant: the same patient, but hemoglobin has dropped to 8.1 g/dL and the surgical team has not yet evaluated the bleed. The system proposes the same action sequence. Layer 1 produces the same epistemically sound output (high VTE risk, moderate bleeding risk). However, Layer 2's rule engine identifies a contraindication: "do not initiate full-dose anticoagulation in the setting of active post-operative bleeding without surgical clearance." The action is flagged as impermissible and routed for mandatory clinician review. This example illustrates agency non-transferability in operation: the diagnostic assessment is identical, but the institutional constraint context renders the same action impermissible.

Layer 3 (Institutional Traceability). Both scenarios are logged: the belief state, proposed action sequence, constraint evaluation results, and final disposition (executed vs. flagged). This enables retrospective audit of whether the governance stack functioned correctly and supports the credentialing evaluation in Phase 2 of the roadmap.

This worked example is intended as a design-validation instrument, tracing a representative clinical trajectory through the full PO-CDP stack to demonstrate how each layer interacts at decision time. Empirical validation via retrospective decision-log analysis is the explicit objective of Phase 1

 doi:10.20944/preprints202605.0723.v1

(Section 5), where the action validity score and override rate operationalize the present framework's claims in a measurable, falsifiable form.

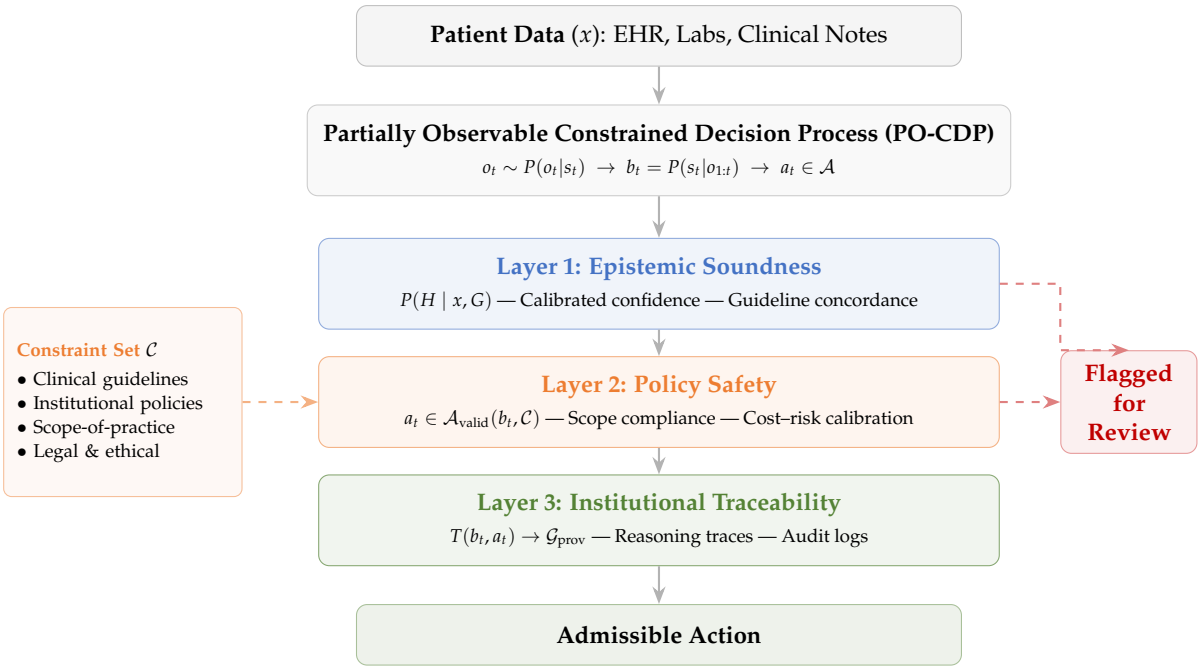

Figure 1. The three-layer governance stack for agentic clinical AI. Patient data enter the partially observable constrained decision process (PO-CDP), generating belief states and action proposals. Layer 1 (epistemic soundness) validates diagnostic outputs against clinical guidelines. Layer 2 (policy safety) filters proposed actions against a multi-dimensional constraint set $\mathcal{C}$ (left), ensuring admissibility under clinical, institutional, scope-of-practice, and legal constraints. Layer 3 (institutional traceability) records full decision provenance for audit and accountability. Actions failing evaluation at Layers 1 or 2 are flagged for mandatory clinician review (right) rather than being silently passed to execution.

4. Compositional Risk in Agentic Clinical Systems

4.1. Beyond Single-Step Errors

The governance stack defines verification guarantees at each layer. A critical question remains: do verified components necessarily compose into a safe system? The answer, in general, is no.

Traditional clinical AI evaluation treats errors as independent and additive. Under this assumption, the total risk of a multi-step clinical pipeline is the sum of individual step risks:

$$R_{\text{total}} = \sum_{i=1}^{n} R_i \tag{4}$$

This additive model holds when errors at each step are independent, such that a misstep in test ordering has no bearing on subsequent treatment selection. In agentic systems, this independence assumption fails; trajectory-level uncertainty accumulates across action steps in ways that single-point confidence estimates cannot capture [6].

4.2. Compositional Risk Formulation

Agentic clinical AI executes multi-step action sequences where each step depends on the outputs of prior steps. The total risk is not merely the sum of individual component risks but includes pairwise interaction terms that capture error amplification across dependent steps. We model this as a supermodular risk function:

$$R_{\text{total}} \geq \sum_{i=1}^{n} R_i + \sum_{(i,j)\in E} \gamma_{ij} \cdot R_i \cdot R_j \tag{5}$$

 doi:10.20944/preprints202605.0723.v1

where E is the edge set of the dependency graph $D_{1..n}$ and $\gamma_{ij} \geq 0$ are pairwise interaction coefficients capturing error amplification between steps i and j. The supermodularity condition ($\partial^2 R_{\text{total}} / \partial R_i \partial R_j \geq 0$) is satisfied by choosing $\gamma_{ij} \geq 0$, ensuring that the marginal risk of an error at any step is non-decreasing in the error rates of its dependencies. When $\gamma_{ij} = 0$ for all pairs, the formulation reduces to the additive model (Eq. 4). When $\gamma_{ij} > 0$, errors compound nonlinearly: a diagnostic hypothesis with 95% confidence ($R_1 = 0.05$), combined with a minor scope-of-practice violation in test ordering ($R_2 = 0.03$), may yield a treatment decision with risk substantially exceeding $R_1 + R_2 = 0.08$ when γ_{12} is large. This result is well established in systems safety engineering [24]: system-level safety is an emergent property that cannot be inferred from component-level safety alone. Even if the epistemic layer is well-calibrated, the policy layer is constraint-compliant, and the traceability layer is fully operational, their interaction can produce unsafe behavior through unanticipated feedback loops, reward hacking [17], or distributional shift across the action pipeline. The inequality in Eq. 5 is a formal consequence of the supermodularity assumption; empirical estimation of its magnitude requires measurement of the γ_{ij} coefficients, methods for which we outline below.

Estimating the interaction coefficients γ_{ij} from observational or prospective data requires study designs that can isolate pairwise error amplification from additive component contributions. Three complementary strategies are available. First, *retrospective regression*: given logs of multi-step AI pipeline decisions and associated patient outcomes, fit a model of observed outcome variance against the additive component-risk baseline (Eq. 4); the residual variance attributable to pairwise step interactions yields γ_{ij} estimates via ordinary least squares or Bayesian regression. Second, *expert elicitation*: structured dependency-mapping workshops using established techniques such as SWIFT (Structured What-If Technique) or bow-tie analysis can elicit prior distributions over γ_{ij} from clinical and systems-safety domain experts, enabling Bayesian updating as empirical data accumulate. Third, *prospective cohort study*: longitudinal recording of complete multi-step decision trajectories with blinded outcome adjudication allows maximum-likelihood estimation of interaction effects in a pre-registered analysis. The Phase 1 action logs and Phase 2 outcome data specified in Figure 2 provide the data infrastructure for all three strategies.

4.3. Implications for Safety Analysis

The compositional risk perspective carries three consequences for governance. First, evaluation must target the *system level*, not merely the component level. Standard benchmarks that evaluate diagnostic accuracy or calibration in isolation are necessary but insufficient; end-to-end evaluation of action sequences against patient outcomes is required. Second, Layer 2 must account for multi-step dependencies, not just single-action admissibility. A policy filter that evaluates each action in isolation may approve a sequence of individually permissible actions whose cumulative effect is unsafe. Third, Layer 3 is not merely an accountability mechanism but a safety mechanism: it enables retrospective analysis of how individually correct components produced adverse system-level outcomes.

4.4. Robustness Under Dependency Uncertainty

A practical limitation of the compositional risk model is that clinical dependencies are often hidden, unrecorded, or dynamically changing. In complex multi-step workflows, the dependency graph $D_{1..n}$ may be partially observable: some step-to-step couplings are documented in care protocols, but others emerge from tacit clinical routines, informal communication patterns, or rare event cascades that do not appear in structured data. This makes empirical estimation of all γ_{ij} coefficients infeasible in early deployment.

The governance stack is designed to remain robust under this limitation through two mechanisms. First, the Layer 3 audit infrastructure provides the data foundation for incremental dependency learning: fine-grained provenance logs capturing the full state of each decision step—including contextual variables, constraint evaluations, and action outcomes—enable retrospective identification of previously unmapped couplings as operational experience accumulates. Second, when γ_{ij} values are unknown or poorly estimated, the constraint satisfaction module can substitute conservative worst-

 doi:10.20944/preprints202605.0723.v1

case upper bounds derived from domain-specific safety envelopes (for example, maximum plausible drug–drug interaction severity from pharmacological databases). Operating under worst-case bounds increases the rate of human escalation, effectively reducing the system's autonomy level until empirical evidence narrows the uncertainty—a form of risk-informed conservatism analogous to the defense-in-depth principle in nuclear safety standards (IEC 61508). Developing adaptive γ_{ij} estimation methods that update continuously from streaming audit logs as operational data accumulate is a priority for Phase 2 implementation.

5. Extended Backcasting Roadmap with Empirical Anchors

The governance stack and compositional risk analysis define *what* must be built. The question of *when and how* these structures should be deployed requires a transition strategy. We extend the backcasting roadmap [4], which identified three temporal pivot points for clinician AI adoption, by operationalizing the transition from the 2030 verifiable AI pivot to the 2035 agentic governance pivot. This gap—five years in which advisory systems must evolve into governed agentic ones—is the critical transition window.

To structure this transition, we define a five-level clinical AI autonomy spectrum (Table 2). This taxonomy is informed by Parasuraman et al.'s framework for automation levels [32] but adapted for the clinical context, where constraints are multi-dimensional (clinical, institutional, legal) rather than purely technical.

Table 2. Clinical AI autonomy levels with operational criteria and governance layer requirements.

Level	Action Scope	Human Role	Layers Req.
L0: Advisory	Generates diagnostic predictions only	Full clinical decision authority	L1
L1: Proposal	Proposes discrete actions; no execution	Mandatory confirmation per action	L1 + L2 (advisory)
L2: Scoped execution	Executes pre-approved actions in narrow domain	Retrospective review of executed actions	L1 + L2 + L3
L3: Supervised autonomy	Executes multi-step sequences in credentialed domain	On-loop monitoring; intervenes on flags	All three
L4: Autonomous management	Manages longitudinal care trajectories	Periodic review; intervenes on drift	All three + calibration

The agency gap becomes a governance concern beginning at L1, where systems first propose actions, and intensifies at L2, where actions are executed without per-step confirmation. Phase boundaries are set to align with institutional readiness timelines: Phase 1 begins in 2027 because current regulatory and infrastructure development cycles (FDA software as a medical device guidance revision, HL7 FHIR policy-as-code pilot programs) suggest a minimum 18-month horizon before institutional deployment is feasible [26]; Phase 2 begins in 2030, coinciding with the 2030 verifiable AI pivot from the original roadmap, which provides the epistemic foundation necessary for credentialing; Phase 3 begins in 2033, requiring at least three years of Phase 2 credentialing data before expanded autonomy can be justified. We define three empirically testable phases, each with measurable milestones (Figure 2).

 doi:10.20944/preprints202605.0723.v1

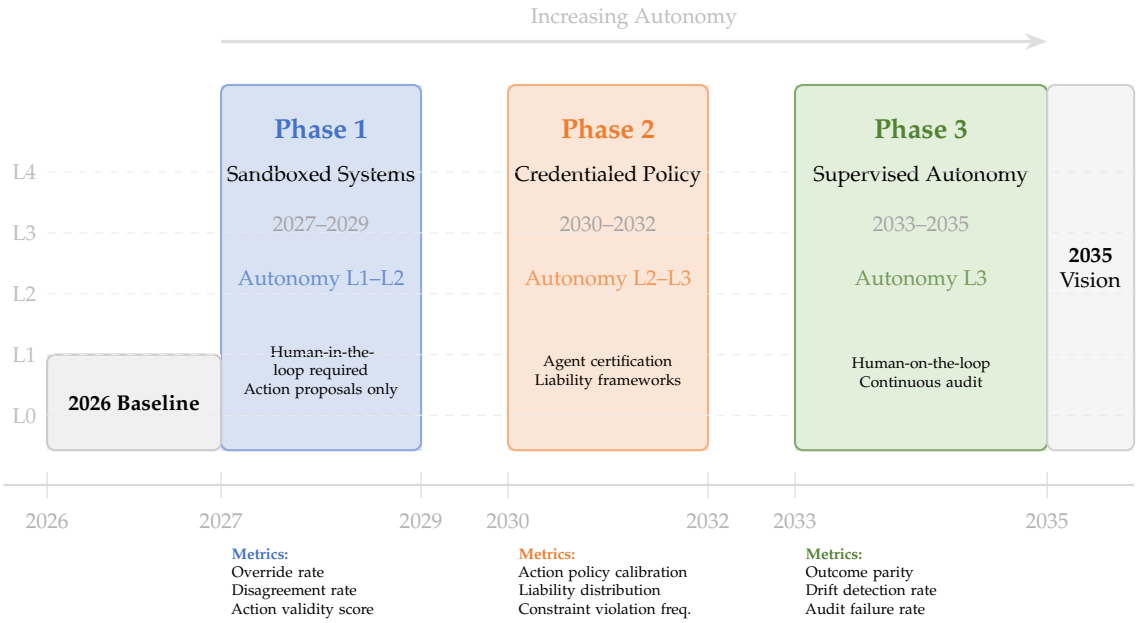

Figure 2. Extended backcasting roadmap for the transition to governed agentic clinical AI (2027–2035). Three empirically testable phases fill the gap between the 2030 verifiable AI pivot and the 2035 agentic governance pivot identified in the original 2040 roadmap [4]. Phase 1 (sandboxed agentic systems) permits action proposals under mandatory human confirmation. Phase 2 (credentialed policy systems) introduces domain-specific agent certification and liability frameworks. Phase 3 (supervised autonomy) enables limited autonomous execution under continuous outcome monitoring. Each phase specifies falsifiable metrics (listed below each block) that determine the pace of transition. Dashed horizontal lines indicate autonomy levels (L0–L4) from the advisory–agentic spectrum.

5.1. Phase 1: Sandboxed Agentic Systems (2027–2029)

The first phase permits limited agentic capability within strictly bounded clinical domains. AI systems may propose action sequences—ordering tests, generating referral recommendations, initiating medication reconciliation—but all proposals require explicit human confirmation before execution. The governance stack operates at partial capacity: Layer 1 (epistemic soundness) is fully operational, building on the dual-process verification architecture; Layer 2 (policy safety) is implemented as a constrained action filter that flags constraint violations but does not autonomously block execution; Layer 3 (traceability) logs all proposed actions and human responses.

The prerequisite for this phase is the institutional deployment of the dual-process verification layer [4]. Without verified diagnostic outputs, agentic proposals lack the epistemic foundation necessary for meaningful constraint evaluation. Empirically testable metrics for this phase include the *override rate* (proportion of proposed actions rejected by clinicians), the *disagreement rate* (frequency of clinician–agent divergence on action appropriateness), and the *action validity score* (proportion of proposed actions that satisfy Layer 2 constraint evaluation). These metrics establish the baseline behavioral profile of agentic systems under human supervision.

5.2. Phase 2: Credentialed Policy Systems (2030–2032)

The second phase introduces formal credentialing of AI agents for specific clinical domains. Credentialing is analogous to board certification but focused on demonstrated calibration performance under constraint: an agent is certified for a specific clinical domain (e.g., post-operative medication management, chronic disease monitoring) if its action proposals satisfy defined accuracy, safety, and constraint-compliance thresholds over a prospective evaluation period.

This phase requires two institutional innovations. First, the development of *agent certification standards*—specialty-specific benchmarks that evaluate not only diagnostic accuracy but policy behavior under constraint, including constraint violation rates, appropriateness of action sequences, and

 doi:10.20944/preprints202605.0723.v1

robustness under distribution shift. These standards would complement emerging regulatory frameworks: the EU AI Act classifies medical AI systems as high-risk, requiring conformity assessments and post-market surveillance [30], while the FDA's predetermined change control plans provide a mechanism for approving adaptive AI systems that evolve within pre-specified bounds [31]. Second, the introduction of *liability frameworks* that recognize AI-assisted decisions as a category of collaborative clinical output, distributing accountability across the clinician, the institution's chief AI officer (CAIO), and the certified governance record [4,23]. Empirically testable metrics include the calibration of action policies (agreement between proposed and expert-adjudicated optimal actions), patterns of liability distribution, and constraint violation frequency across credentialed domains.

5.3. Phase 3: Supervised Autonomy (2033–2035)

The third phase enables limited autonomous execution in defined clinical domains. Agents with demonstrated calibration track records—established during Phases 1 and 2—are granted expanded scope, contingent on continuous outcome-based performance monitoring. The CAIO conducts quarterly calibration audits; agents that drift below performance thresholds have their autonomy level reduced or revoked [4].

This phase operationalizes the 2035 agentic governance pivot from the original roadmap. Its distinguishing feature is the shift from human-in-the-loop confirmation (Phase 1) to human-on-the-loop oversight: the clinician does not approve each action but monitors aggregate system behavior through the traceability layer and intervenes when audit flags are raised. Empirically testable metrics include outcome parity between AI-managed and clinician-managed cases within credentialed domains, drift detection rates from continuous calibration monitoring, and audit failure rates from the traceability layer.

A critical feature of this phased approach is that each phase is *falsifiable*. If Phase 1 metrics show persistently high override rates or constraint violation frequencies, the transition to Phase 2 is delayed. If credentialed agents in Phase 2 exhibit calibration drift or unexpected compositional failures, the scope of autonomous execution in Phase 3 is restricted. The roadmap does not assume linear progress; it specifies adaptive checkpoints at which the evidence determines the pace of transition.

Phase transitions are treated as formal decision points requiring pre-registered statistical evaluation, not qualitative assessments. Table 3 provides operational definitions and transition thresholds for the three representative metrics named in Figure 2; formal definitions for all nine phase-specific metrics will be specified in a pre-registered analysis plan prior to Phase 1 deployment.

Table 3. Formal statistical definitions and phase-transition thresholds for representative roadmap metrics.

Metric	Operational Definition	Statistical Method	Transition Threshold
Action validity score (Phase 1)	Proportion of proposed actions passing Layer 2 constraint evaluation within a 90-day rolling window	Two-sided 95% Wilson score confidence interval for binomial proportion	≥ 0.95; lower CI bound > 0.90
Outcome parity (Phase 3)	Standardized risk difference in primary adverse event rate between AI-managed and clinician-managed cases in a prospective comparative cohort	Two-sided equivalence test ($\alpha = 0.05$); powered at 80%; Bonferroni correction for co-primary endpoints	Equivalence margin $\lvert\Delta\rvert \leq 0.05$ on adverse event rate
Drift detection rate (Phase 3)	Alert frequency from CUSUM control chart applied to rolling Layer 1 calibration metrics	Cumulative sum chart with 3σ alert threshold; average run length (ARL) ≥ 500 under null	Alert rate below pre-specified operational threshold

 doi:10.20944/preprints202605.0723.v1

6. Critical Tensions and Open Problems

6.1. The Calibration–Drift Paradox

The PO-CDP framework assumes a stable constraint set $\mathcal{C}$ and a stationary belief distribution. In practice, both drift. As agents learn from local institutional data, their calibration profiles diverge from the original verification benchmarks. The 2026 backcasting analysis identified this as an equity risk—locally deployed SLMs may produce well-calibrated outputs for dominant populations and systematically miscalibrated outputs for underrepresented groups [4,21]. For agentic systems, the risk is compounded: drift occurs not only in diagnostic calibration but in *action patterns*. An agent trained primarily on data from a tertiary academic medical center may propose action sequences that are appropriate in that context but unsafe or unavailable in a community hospital setting. Federated calibration approaches [22] offer a partial solution for diagnostic models, but mechanisms for detecting and correcting action-level drift remain an open research problem.

6.2. From Automation Bias to Audit Burden Shift

The automation bias literature documents a well-established risk: as systems become more reliable, clinicians may accept AI outputs without appropriate critical evaluation [20]. For agentic systems, this risk escalates. When an agent autonomously manages medication reconciliation, the clinician's opportunity to catch errors shifts from the *decision point* to the *audit point*, a fundamentally different cognitive task requiring retrospective reasoning about actions the clinician did not directly observe. This is a shift from real-time vigilance to retrospective auditing, a skill not currently taught in medical education. The implication for Phase 3 is that supervised autonomy requires clinician training in audit reasoning, connecting to the 2040 futures literacy pivot [4].

6.3. Interoperability as a Hard Constraint

The policy safety layer (Layer 2) requires computable institutional policies: machine-readable representations of formulary constraints, scope-of-practice rules, and care pathway definitions. Current EHR systems, built on HL7/FHIR interoperability standards, do not uniformly support this capability. Without *policy-as-code*, structured, machine-interpretable representations of clinical and institutional constraints, the action validation layer cannot function. This is not a regulatory or governance barrier but an infrastructure prerequisite. The development of policy-as-code standards for health care is a necessary precondition for Phase 2 and represents a distinct technical workstream that must proceed in parallel with governance design.

6.4. Equity and Infrastructure Asymmetry

The 2026 backcasting analysis acknowledged its US-centric scope limitation and the resource disparities between well-resourced and safety-net health systems [4]. For agentic governance, these disparities intensify. Rural and safety-net systems may lack the compute infrastructure required for local SLM deployment, making the CAIO-as-a-service model a structural prerequisite rather than an optional alternative for a substantial portion of the health care system. Jurisdictional differences in medical liability make cross-border credentialing frameworks complex. Agentic systems trained primarily on data from well-resourced institutions may produce action sequences that are inappropriate in resource-limited settings. A multi-jurisdictional extension of the roadmap, accounting for varying regulatory, infrastructural, and demographic contexts, is a necessary parallel research agenda. The WHO's guidance on AI governance for health provides a foundation for such cross-jurisdictional harmonization but does not yet address agentic systems specifically [34].

6.5. Patient Agency and Informed Consent

This analysis has been clinician-centric, consistent with the governance stack's focus on institutional and professional accountability. However, the transition to agentic AI—where systems act autonomously rather than merely advise—has direct implications for patient autonomy and informed

 doi:10.20944/preprints202605.0723.v1

consent. When an AI agent autonomously manages medication reconciliation or initiates referral pathways, the patient's understanding of who is making clinical decisions and on what basis becomes materially different from the traditional physician–patient model. Patients may not be aware that an agentic system, rather than their physician, proposed a specific intervention. The governance stack's traceability layer (Layer 3) could support consent transparency—providing patients with accessible records of AI-initiated actions—but the design of patient-facing governance mechanisms remains an open problem that this paper does not address and that warrants dedicated investigation.

7. Institutional Design and Human Factors

7.1. Governance Beyond Technical Architecture

The three-layer governance stack is a sociotechnical system, not a purely technical one. Its effective deployment requires institutional conditions that cannot be specified in software: workflow redesign to accommodate human-on-the-loop oversight, cultural adaptation to distributed clinical agency, and organizational commitment to the audit infrastructure that Layer 3 demands. The automation bias and audit burden analysis in Section 6 demonstrates that technical safeguards alone are insufficient; the human operator's cognitive relationship to the system must be intentionally designed.

Evidence from other safety-critical domains supports this view. Bainbridge's classic analysis of automation in process control identified the "ironies of automation": designers automate the tasks that are easiest to formalize, leaving operators with the most difficult and cognitively demanding monitoring tasks for which they are the least practiced [33]. In aviation, the transition through automation levels revealed that intermediate autonomy (analogous to our L1–L2) created the highest accident risk, because pilots were removed from active control yet required to intervene during emergencies—a pattern directly paralleling the clinician's position in Phase 1 of our roadmap [32]. Nuclear power plant operations demonstrated that effective human-on-the-loop oversight requires dedicated training in retrospective systems reasoning, a competency distinct from the prospective decision-making emphasized in operator training [33]. These precedents suggest that the governance stack's success depends not only on its technical architecture but on institutional investment in new oversight competencies.

7.2. New Clinical Competencies

The transition to agentic clinical AI requires competencies that are not currently part of medical education. Audit reasoning—the ability to retrospectively evaluate the appropriateness of AI-generated action sequences—is distinct from the prospective clinical reasoning taught in current curricula. Oversight of autonomous systems requires understanding the governance stack's failure modes (Table 1) and the conditions under which human intervention is necessary. These competencies extend the futures literacy framework proposed as the 2040 pivot [4], connecting foresight methods to the practical demands of governed human-AI teaming.

7.3. Institutional Role Evolution

The CAIO role, proposed in the 2026 roadmap as the institutional innovation bridging clinical and technical governance, must expand to encompass agentic oversight. The CAIO's mandate should include certification of agent scope, quarterly audit of action compliance, escalation protocols for out-of-scope actions, and equity reporting across demographic strata. The emergence of dedicated AI governance boards within health systems—analogous to institutional review boards for research ethics—may provide the collective decision-making infrastructure that individual CAIOs cannot sustain alone.

 doi:10.20944/preprints202605.0723.v1

8. Discussion

The central argument of this paper is that the transition to agentic clinical AI is not a scaling problem amenable to incremental technical improvement. It is a shift from prediction correctness to policy safety under constraint. Three consequences follow.

First, verification must move from truth to permissibility. The question is no longer whether the system's diagnosis is correct, but whether its proposed actions are allowed, appropriate, and safe given the patient's context and the institution's constraints. This requires mechanisms, notably the action validation layer, that have no precedent in current clinical AI systems.

Second, safety is a systems property, not a model property. The compositional risk analysis formally predicts that individually verified components can produce unsafe system-level behavior through nonlinear error interactions. This finding aligns with established results in systems safety engineering [24] and challenges the prevailing model-centric evaluation paradigm in clinical AI.

Third, institutional design requirements are as consequential as the technical ones. Layer 3, institutional traceability, is not an afterthought but a constitutive element of safe agentic operation. Without auditable provenance graphs, distributed clinical agency collapses into unassignable responsibility, and the liability frameworks necessary for Phase 2 and Phase 3 become legally incoherent.

Together, these three consequences redefine the research agenda for clinical AI: from improving model accuracy to designing, implementing, and validating governance systems that enable safe machine agency under the constraints of clinical practice.

9. Conclusions

The agency gap, the structural mismatch between validated predictions and unvalidated action policies, is the central governance challenge as clinical AI moves toward autonomous action. The verification problem shifts from whether the system's beliefs are correct to whether its actions are permissible under clinical, institutional, and legal constraints.

We have formalized this shift using a PO-CDP framework, demonstrated the non-transferability of trust from prediction to action, proposed a three-layer governance stack with verifiable guarantees, and extended the backcasting roadmap with empirically testable milestones. The compositional risk analysis shows that safe agentic clinical AI requires system-level evaluation, not merely component-level validation.

Three research and policy priorities emerge. First, the development of action-level validation benchmarks: standardized evaluation frameworks for policy behavior under constraint, analogous to the diagnostic accuracy benchmarks that currently govern clinical AI evaluation. Second, the design of agent credentialing frameworks: institutional mechanisms for certifying that AI agents satisfy defined safety and performance thresholds within specific clinical domains. Third, the construction of governance infrastructure: the institutional, legal, and educational structures, including expanded CAIO mandates, liability frameworks for distributed clinical agency, and audit reasoning curricula, necessary to support the transition from supervised to autonomous agentic operation.

The future of clinical AI will depend not on model accuracy alone but on the capacity to govern machine agency under the constraints of clinical practice.

Funding: The author received no external funding for this research.

Conflicts of Interest: The author declares no conflicts of interest.

Data Availability Statement: No datasets were generated or analyzed during the current study. This paper presents a theoretical governance framework; all illustrative examples are hypothetical.

Informed Consent Statement: This study does not involve human participants, patient data, or biological material. Institutional review board approval was not required.

 doi:10.20944/preprints202605.0723.v1

References

1. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. *Nat Med*. 2019;25(1):44–56. doi:10.1038/s41591-018-0300-7. https://www.nature.com/articles/s41591-018-0300-7
2. Rajpurkar P, Chen E, Banerjee O, Topol EJ. AI in health and medicine. *Nat Med*. 2022;28(1):31–38. doi:10.1038/s41591-021-01614-0. https://www.nature.com/articles/s41591-021-01614-0
3. Yu Y, Gomez-Cabello CA, Haider SA, et al. Enhancing clinician trust in AI diagnostics: a dynamic framework for confidence calibration and transparency. *Diagnostics*. 2025;15(17):2204. doi:10.3390/diagnostics15172204. https://www.mdpi.com/2075-4418/15/17/2204
4. Yu Y. Backcasting the trust gap: a strategic road map for clinician adoption of AI diagnostics by 2040. *J Med Internet Res*. 2026;28:e94234. doi:10.2196/94234. https://www.jmir.org/2026/1/e94234
5. Yu Y. Agentic AI in healthcare: bridging the gap between computational promise and clinical evidence. *Research Square*. Preprint posted online April 14, 2026. doi:10.21203/rs.3.rs-9374197/v1. https://www.researchsquare.com/article/rs-9374197/v1
6. Yu Y. Variability-aware trust in agentic clinical AI: modeling and measuring trajectory-level uncertainty. *Research Square*. Preprint posted online April 7, 2026. doi:10.21203/rs.3.rs-9326470/v1. https://www.researchsquare.com/article/rs-9326470/v1
7. Yu Y. Causal-constrained reinforcement learning for revenue cycle optimization. *Research Square*. Preprint posted online April 21, 2026. doi:10.21203/rs.3.rs-9465761/v1. https://www.researchsquare.com/article/rs-9465761/v1
8. Yu Y. Hybrid-code: a privacy-preserving, redundant multi-agent framework for reliable local clinical coding. *arXiv*. Preprint posted online December 26, 2025. arXiv:2512.23743. https://arxiv.org/abs/2512.23743
9. Yu Y, Gomez-Cabello CA, Makarova S, Parte Y, Borna S, Haider SA, Genovese A, Prabha S, Forte AJ. Using large language models to retrieve critical data from clinical processes and business rules. *Bioengineering*. 2025;12(1):17. doi:10.3390/bioengineering12010017. https://www.mdpi.com/2306-5354/12/1/17
10. Moor M, Banerjee O, Abad ZSH, et al. Foundation models for generalist medical artificial intelligence. *Nature*. 2023;616(7956):259–265. doi:10.1038/s41586-023-05881-4. https://www.nature.com/articles/s41586-023-05881-4
11. Amann J, Blasimme A, Vayena E, Frey D, Madai VI, Precise4Q consortium. Explainability for artificial intelligence in healthcare: a multidisciplinary perspective. *BMC Med Inform Decis Mak*. 2020;20(1):310. doi:10.1186/s12911-020-01332-6. https://bmcmedinformdecismak.biomedcentral.com/articles/10.1186/s12911-020-01332-6
12. Singhal K, Azizi S, Tu T, et al. Large language models encode clinical knowledge. *Nature*. 2023;620(7972):172–180. doi:10.1038/s41586-023-06291-2. https://www.nature.com/articles/s41586-023-06291-2
13. Puterman ML. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons; 2014. doi:10.1002/9780470316887. https://onlinelibrary.wiley.com/doi/book/10.1002/9780470316887
14. Altman E. *Constrained Markov Decision Processes*. Chapman & Hall/CRC; 1999. doi:10.1201/9781420035736. https://www.routledge.com/Constrained-Markov-Decision-Processes/Altman/p/book/9780849303820
15. Kaelbling LP, Littman ML, Cassandra AR. Planning and acting in partially observable stochastic domains. *Artif Intell*. 1998;101(1–2):99–134. doi:10.1016/S0004-3702(98)00023-X. https://www.sciencedirect.com/science/article/pii/S000437029800023X
16. Guo C, Pleiss G, Sun Y, Weinberger KQ. On calibration of modern neural networks. In: *Proceedings of the 34th International Conference on Machine Learning*. 2017:1321–1330. Available at: https://proceedings.mlr.press/v70/guo17a.html
17. Amodei D, Olah C, Steinhardt J, Christiano P, Schulman J, Mané D. Concrete problems in AI safety. *arXiv*. Preprint posted online June 21, 2016. arXiv:1606.06565. doi:10.48550/arXiv.1606.06565. https://arxiv.org/abs/1606.06565
18. García J, Fernández F. A comprehensive survey on safe reinforcement learning. *J Mach Learn Res*. 2015;16(1):1437–1480. Available at: https://jmlr.org/papers/v16/garcia15a.html
19. Seto D, Krogh B, Sha L, Chutinan A. The simplex architecture for safe online control system upgrades. In: *Proceedings of the 1998 American Control Conference*. 1998:3504–3508. doi:10.1109/ACC.1998.703296. https://ieeexplore.ieee.org/document/703296
20. Goddard K, Roudsari A, Wyatt JC. Automation bias: a systematic review of frequency, effect mediators, and mitigators. *J Am Med Inform Assoc*. 2012;19(1):121–127. doi:10.1136/amiajnl-2011-000089. https://academic.oup.com/jamia/article/19/1/121/730102

 doi:10.20944/preprints202605.0723.v1

21. Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*. 2019;366(6464):447–453. doi:10.1126/science.aax2342. https://www.science.org/doi/10.1126/science.aax2342
22. Yu Y. AdaptiveFedLoRA: drift-aware adaptive LoRA rank scheduling for federated medical small language models. *medRxiv*. Preprint posted online January 21, 2026. doi:10.64898/2026.01.18.26344237. https://www.medrxiv.org/content/10.64898/2026.01.18.26344237v1
23. Price WN II, Gerke S, Cohen IG. Potential liability for physicians using artificial intelligence. *JAMA*. 2019;322(18):1765–1766. doi:10.1001/jama.2019.15064. https://jamanetwork.com/journals/jama/fullarticle/2753070
24. Leveson N. *Engineering a Safer World: Systems Thinking Applied to Safety*. MIT Press; 2011. doi:10.7551/mitpress/8178.001.0001. https://mitpress.mit.edu/9780262016629/engineering-a-safer-world/
25. Collaco BG, Haider SA, Prabha S, et al. The role of agentic artificial intelligence in healthcare: a scoping review. *npj Digit Med*. 2026;9:345. doi:10.1038/s41746-026-02517-5. https://www.nature.com/articles/s41746-026-02517-5
26. US Food and Drug Administration. Artificial intelligence/machine learning (AI/ML)-based software as a medical device (SaMD) action plan. 2021. https://www.fda.gov/media/145022/download
27. Bai Y, Kadavath S, Kundu S, et al. Constitutional AI: harmlessness from AI feedback. *arXiv*. Preprint posted online December 15, 2022. arXiv:2212.08073. https://arxiv.org/abs/2212.08073
28. Ouyang L, Wu J, Jiang X, et al. Training language models to follow instructions with human feedback. *Adv Neural Inf Process Syst*. 2022;35:27730–27744. arXiv:2203.02155. https://arxiv.org/abs/2203.02155
29. Rebedea T, Dinu R, Sreedhar MN, Parisien C, Cohen J. NeMo guardrails: a toolkit for controllable and safe LLM applications with programmable rails. In: *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. 2023:431–445. doi:10.18653/v1/2023.emnlp-demo.40. https://aclanthology.org/2023.emnlp-demo.40
30. European Parliament and Council. Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). *Off J Eur Union*. 2024;L 2024/1689. https://eur-lex.europa.eu/eli/reg/2024/1689
31. US Food and Drug Administration. Marketing submission recommendations for a predetermined change control plan for artificial intelligence/machine learning (AI/ML)-enabled device software functions. Guidance for industry and FDA staff. 2023. https://www.fda.gov/media/166764/download
32. Parasuraman R, Sheridan TB, Wickens CD. A model for types and levels of human interaction with automation. *IEEE Trans Syst Man Cybern Part A Syst Hum*. 2000;30(3):286–297. doi:10.1109/3468.844354. https://ieeexplore.ieee.org/document/844354
33. Bainbridge L. Ironies of automation. *Automatica*. 1983;19(6):775–779. doi:10.1016/0005-1098(83)90046-8. https://www.sciencedirect.com/science/article/pii/0005109883900468
34. World Health Organization. Ethics and governance of artificial intelligence for health: WHO guidance. 2021. https://www.who.int/publications/i/item/9789240029200