Viewpoint

Recoverability-Centered Trajectory Governance for Long-Horizon Clinical AI Agents

Jialin Liu[1,2*], MD; Siru Liu[3*], PhD

[1]Information Center, West China Hospital of Sichuan University, Chengdu, Sichuan, China
[2]Department of Otolaryngology-Head and Neck Surgery, West China Hospital of Sichuan University, Chengdu, Sichuan, China
[3]Department of Biomedical Informatics, Vanderbilt University Medical Center, Nashville, TN, United States
*all authors contributed equally

Corresponding Author:

Siru Liu, PhD
Department of Biomedical Informatics
Vanderbilt University Medical Center
2525 West End Ave, Suite 1475
Nashville, TN 37212
United States
Phone: 1 615 875 5216
Email: siru.liu@vumc.org

Abstract

Clinical AI is moving from systems that generate discrete predictions and recommendations toward agents capable of pursuing clinical objectives through repeated tool use, communication, planning, and state-changing actions. Emerging systems support multistep electronic health record workflows, and increasingly envision longitudinal care across repeated encounters. This transition expands safety assessment beyond discrete outputs and actions. Through its actions, an agent can help create the clinical state on which its subsequent decisions depend. Consequently, individually defensible actions may compose into an unsafe trajectory, even when no single action is clearly erroneous. We propose recoverability-centered trajectory governance as a clinical synthesis of established safety approaches, linking patient-specific corrective capacity to the continuing justification for delegated authority. Its central objective is clinical recoverability: preserving the opportunity for the human-AI care system to redirect care before clinically acceptable corrective pathways are exhausted. We define recovery margin as the remaining opportunity for such correction, and reauthorization boundaries as conditions requiring delegated authority to end or narrow pending renewed human judgment. This framing distinguishes reversibility, primarily a property of an action, from recoverability, a property of an evolving clinical trajectory. The operational framework combines persistent obligation tracking, independent monitoring, explicit reauthorization, authority contraction, graceful degradation, independent escalation, and verified human takeover. Oversight should concentrate at consequential trajectory junctions. Evaluation should combine trajectory stress testing, component-specific recovery-margin indicators, interruption-and-takeover testing, and population-level surveillance. Trajectory governance complements product lifecycle oversight by examining whether continued patient-specific delegation remains justified as care evolves. Because corrective capacity depends partly on access to timely monitoring and human rescue, recoverability is also an equity issue. The defining safety question is therefore not only whether an action is acceptable but whether continued delegation remains justified. The proposed indicators and governance mechanisms require domain-specific operationalization and empirical validation, including comparison with established safeguards.

J Med Internet Res 2026;28:e110563; doi: 10.2196/110563

Keywords: AI; agentic AI; patient safety; clinical decision support; trajectory safety; recoverability; human oversight; AI governance; risk management; longitudinal care

Introduction

Clinical AI is moving from generating bounded outputs toward agents that can participate in changing clinical states through planning, reasoning, tool use, communication, and action [1-3]. Recent systems are also moving toward longitudinal management [4]. In a randomized, blinded virtual study, AMIE (Articulate Medical Intelligence

Explorer) revised management plans across simulated multivisit scenarios and demonstrated overall management reasoning noninferior to that of primary care physicians [4]. Yet longitudinal capability should not be equated with reliable execution. PhysicianBench evaluated agents on 100 long-horizon electronic health record (EHR) tasks; the best-performing system succeeded on 46% of tasks [5]. These studies assess distinct capabilities. One evaluates management reasoning across visits, whereas the other evaluates end-to-end execution of multistep EHR workflows. Their results are not directly comparable. Real-world deployment remains early, with most agents handling administrative work, although an evidence map documents applications in diagnosis and clinical management [6].

Long-horizon autonomy could support scalable, personalized care, but its safety assessment must address cumulative patient-specific consequences. The 2025 draft guidance from the US Food and Drug Administration (FDA) for AI-enabled device software functions and the European Union AI Act incorporate lifecycle risk management, postmarket monitoring, and mechanisms for managing system changes [7,8]. These frameworks primarily organize oversight around the AI system and its lifecycle; the proposed framework makes the continuing justification for delegation within an evolving patient trajectory explicit. Clinical AI evaluation presents a related challenge. MedQA assesses pointwise accuracy [9], while HealthBench [10,11] and PhysicianBench [5] extend evaluation to multiturn conversations and multistep EHR tasks. These evaluations nevertheless remain bounded by predefined interactions or task environments and are not designed to establish safety across open-ended trajectories in which prior actions carry across episodes of care, shape subsequent states, create obligations, and constrain corrective options.

Building on systems-theoretic safety engineering and treatment-strategy research [12,13], we argue that safety for long-horizon clinical AI agents does not necessarily compose because individually appropriate actions can become unsafe as their effects interact or accumulate. Delegated actions may create obligations, narrow corrective options, and consume recovery margin before harm becomes apparent. We therefore propose recoverability-centered trajectory governance, linking patient-specific corrective capacity to the continuing justification for delegated authority. We first characterize long-horizon exposure and trajectory-level failure, explain the limitations of pointwise evaluation alone, and define clinical recoverability and recovery margin. We next clarify the contribution relative to existing approaches and develop the operational framework through obligation tracking, monitoring, reauthorization, authority contraction, graceful degradation, independent escalation, and verified human takeover, illustrating these mechanisms through the anticoagulation example. Finally, we present evaluation methods, broader governance implications, limitations, and a staged empirical validation pathway. The governing principle is to reassess delegation as corrective capacity or supporting conditions change, restricting authority while clinically acceptable correction remains feasible and preserving essential monitoring and rescue functions during transition.

From Long-Horizon Exposure to Clinical Recoverability

Long-Horizon Exposure

Long-horizon clinical agency is not defined by elapsed time alone. We conceptualize long-horizon exposure in terms of 3 interacting dimensions: (1) temporal extension, the duration over which clinically relevant consequences remain unresolved; (2) action depth, the number of consequential agent actions taken since the last clinician reassessment or reauthorization; and (3) causal coupling, the extent to which earlier agent actions alter the information, workflows, clinical states, or options shaping subsequent decisions. These dimensions characterize the deployed model-task-workflow configuration rather than the model alone [14]. AMIE primarily illustrates temporal extension across simulated visits [4], whereas PhysicianBench emphasizes action depth and within-task causal coupling through multistep EHR workflows [5]. Even low action depth can produce strong causal coupling: a single summary stored in the EHR may influence later clinicians or agents, while earlier actions may create obligations or downstream consequences that carry across episodes of care [15-18].

Trajectory-Level Failure From Locally Acceptable Actions

Earlier actions can reshape the conditions governing later decisions, so individually acceptable actions need not compose into a safe trajectory. System safety research describes related hazards through boundary migration [19], tight coupling [20], drift into failure [21], and clinical cascades [22]. The following illustrative warfarin scenario shows how such a failure could arise in long-horizon clinical AI.

Consider a patient receiving maintenance warfarin for venous thromboembolism, with previously stable international normalized ratio (INR) control, who is traveling when monitoring is due. Under a clinician-approved protocol, an agent briefly reschedules the test, continues the authorized regimen, reports that the last measured INR was therapeutic, and closes the completed outreach task. Each action may be locally defensible provided rescheduling eligibility and continued monitoring are preserved; a recent therapeutic INR alone does not establish eligibility. Warfarin monitoring must also account for interacting medicines and changes in clinical status [23]. A composition failure can arise if task closure removes the only reminder, the deferred test lacks an accountable owner, or reassurance is mistaken for evidence of current stability. The monitoring obligation persists while the mechanisms supporting timely completion weaken. This illustrates how permitted actions and incomplete coordination can produce an unsafe trajectory without an initial diagnostic error.

Current benchmarks capture aspects of this problem. CHI-Bench includes stateful workflow constraints [24], while PatientAgentBench reports task-completion rates of 93% to 100% alongside triage-quality performance of 32% to 88% [25]. However, these benchmarks do not directly test persistent cross-episode carryover of agent-created obligations and constraints. Task completion alone, therefore, does not establish trajectory safety.

Clinical Recoverability and the Recovery Margin

Clinical recoverability is the capacity of the human-AI care system, at a given point in an agent-mediated trajectory, to perform 4 functions before corrective pathways are exhausted: recognize that continued delegated action has become unsafe or inadequately authorized; interrupt such action; reconstruct the relevant trajectory state; and redirect care toward a clinically acceptable course. Clinical acceptability refers to a course judged clinically appropriate by the responsible clinician or care team and consistent with the patient's goals and preferences where these are known or can be elicited.

We use the term recovery margin to denote the remaining opportunity for such correction. This margin has at least 3 components: biological reserve, whether patient physiology and the time-dependent therapeutic window still permit effective correction; operational reserve, whether monitoring, human review, patient contact, and rescue can occur within the clinically required time; and epistemic reserve, whether sufficient reliable information remains available to recognize the need for correction and determine an appropriate response. Biological reserve is a property of the patient rather than an agent-controlled resource; however, agent actions, omissions, or delays may nevertheless deplete the physiological opportunity for successful correction. These dimensions reflect familiar clinical concerns, including time-sensitive rescue, situation awareness, and information integrity [15,16,26], but require domain-specific operationalization and validation. They should not be assumed to form a single universally applicable scalar measure.

This framing draws on failure-to-rescue research [27], resilience engineering [28], and deterioration-response systems [26], together with AI-safety concepts of irreversible state change and loss of future options.

In this framework, reversibility concerns whether an action can be undone, whereas clinical recoverability concerns whether the evolving care trajectory can still be redirected toward a clinically acceptable course. An irreversible action may leave a recoverable trajectory if compensatory options and monitoring remain available; conversely, a technically reversible action may consume recovery margin if its consequences are difficult to detect, responsibility becomes ambiguous, or timely reassessment is unavailable. The governing question is therefore whether safe redirection remains feasible given the accumulated clinical state, outstanding obligations, available corrective pathways, and time remaining. To preserve clinical recoverability, human intervention must occur while sufficient recovery margin remains.

Operationalizing Recovery Margin

For maintenance-warfarin monitoring, we propose candidate indicators for domain-specific validation. Biological indicators include the latest INR and its trend when available, bleeding or thromboembolic symptoms, and clinical changes affecting anticoagulation [23,29]. Monitoring must account for interacting medicines and changes in clinical status [23]. These indicators inform risk but do not directly measure the time remaining for successful rescue. Operational indicators include time remaining until the clinically approved test deadline, estimated time to obtain and review a result, patient reachability, and reviewer availability. Epistemic indicators include INR age, completeness of medication reconciliation, provenance of dose instructions, and confirmation that the result belongs to the correct patient.

A missing current INR represents uncertainty, not evidence of preserved biological reserve. Assessment should retain visibility of each component, and favorable values in one should not compensate for a safety-critical deficit in another. The thresholds linking these indicators to reauthorization or authority contraction require clinical specification and empirical validation; they are proposed design elements, not validated clinical rules.

Contribution Relative to Existing Safety Approaches

Recoverability-centered trajectory governance is a proposed clinical synthesis of established safety approaches, including agentic AI safety, resilience engineering, automation, and clinical handoff. Its proposed contribution is to link patient-specific corrective capacity to the continuing justification for delegated authority (Table 1).

Table 1. Relationship to existing safety approaches[a].

Existing approach	Established contribution	Proposed clinical specification
Runtime governance and assurance	Constrain execution and adjust control as operating conditions change [30]	Reassess continued delegation using recovery margin, outstanding obligations, and available corrective pathways
Human oversight and adaptive automation	Allocate control and address limitations of human intervention [31-35]	Use declining recovery margin or materially changed delegation conditions to trigger reassessment and authority contraction, accounting for actual takeover capacity

Existing approach	Established contribution	Proposed clinical specification
Graceful degradation	Preserve essential functions as capabilities decline [36]	Preserve monitoring, obligation tracking, trajectory reconstruction, uncertainty communication, and independent escalation during restricted operation
Resilience engineering and failure-to-rescue research	Examine adaptive capacity and the conditions enabling timely rescue [26-28]	Assess the biological, operational, and epistemic capacity remaining for clinically acceptable correction
Clinical handoff	Support continuity of care and transfer of responsibility [37]	Transfer consequential agent history and unresolved obligations; verify trajectory reconstruction and acceptance of responsibility
Safety and assurance cases, including dynamic approaches	Link claims to evidence and support reassessment as conditions change [38,39]	Document the assumptions, evidence, obligations, and recovery arrangements supporting continued clinical delegation

[a]The proposed contribution lies in organizing these mechanisms around clinical recoverability as a condition for continued delegation. Whether this synthesis improves safety beyond established safeguards requires domain-specific operationalization and empirical comparison.

Recoverability-Centered Trajectory Governance

Obligations and Option Foreclosure

Consequential clinical actions can create obligations and foreclose options. Medication changes or deferred investigations may require subsequent monitoring or reassessment, while closing tasks, canceling referrals, or declining escalation may weaken pathways for detecting and correcting deterioration. Outstanding obligations increase what must still be done; option foreclosure reduces what can still be done. Both can consume recovery margin.

Option loss need not follow an obvious error. Work on preserving future options and system safety provides relevant foundations [40-42]. Routine task closure may remove monitoring opportunities, and reassurance may alter help-seeking, even if later corrected. Biological capacity for rescue may remain while opportunities for timely detection diminish.

Trajectory governance must also capture omissions, such as incomplete testing, referrals, or result follow-up [17,18], which may generate no discrete event detectable by action-based monitoring. We propose an obligation ledger recording the required action, deadline, and accountable owner. Obligations remain open until verified fulfillment or authorized cancellation with a recorded clinical rationale. Transfer requires the receiving owner's acceptance and does not close the obligation. Records should persist across model updates, agent replacement, and handoffs.

A complementary register should track active monitoring, reassessment, referral, escalation, and rescue pathways. Administrative completion must remain distinct from clinical fulfillment. In the warfarin example, closing outreach leaves the rescheduled INR obligation open, with a deadline, a clinical owner, a test location, and a fallback contact route. Fulfillment requires a result matched to the patient and the obligation, documented review, and completion or accountable assignment of required follow-up. Any newly assigned obligations remain open.

Monitoring and Verification Architecture

The agent should not be the sole source of safety-critical detection or the sole authority for closing obligations. Building on runtime-assurance architectures that separate safety supervision from the primary controller [30], we propose a supervisory service outside the agent's planning loop, with a designated accountable owner under health-system governance, to reconcile timestamped EHR events, laboratory data, communication acknowledgments, and ledger states. Deterministic rules can check deadlines, ownership, predefined prohibited actions, and failed contact; ambiguous findings require clinical review, with any model assistance validated for its intended use and documented in the trajectory safety case. Closure requires specified completion evidence; absent, stale, conflicting, or mismatched evidence should prevent closure pending resolution.

An execution gate should check consequential tool calls against current permissions, and time-based monitoring should detect overdue obligations even when the agent takes no further action. Each rule activation should record its trigger, version, rationale, and resulting restriction. Access controls should prevent the agent from modifying supervisory thresholds, completion criteria, its own permissions, or underlying verification evidence.

Architectural separation does not eliminate shared failures. Heartbeat checks, missing-event reconciliation, and backup routes outside the failed workflow require explicit testing, including failures of the primary agent, the supervisory service, the communication route, and shared infrastructure. Verification should distinguish detection of a safety-critical condition, enforcement of the required restriction, and confirmed delivery of escalation to an accountable recipient. These functions should not depend exclusively on the primary agent. Drawing on structured clinical handoff practices [37], we propose that takeover requires confirmation of the relevant trajectory state and acceptance of responsibility; acknowledgment of an alert alone is insufficient. Agent self-escalation is an additional signal, not evidence that these safeguards are effective.

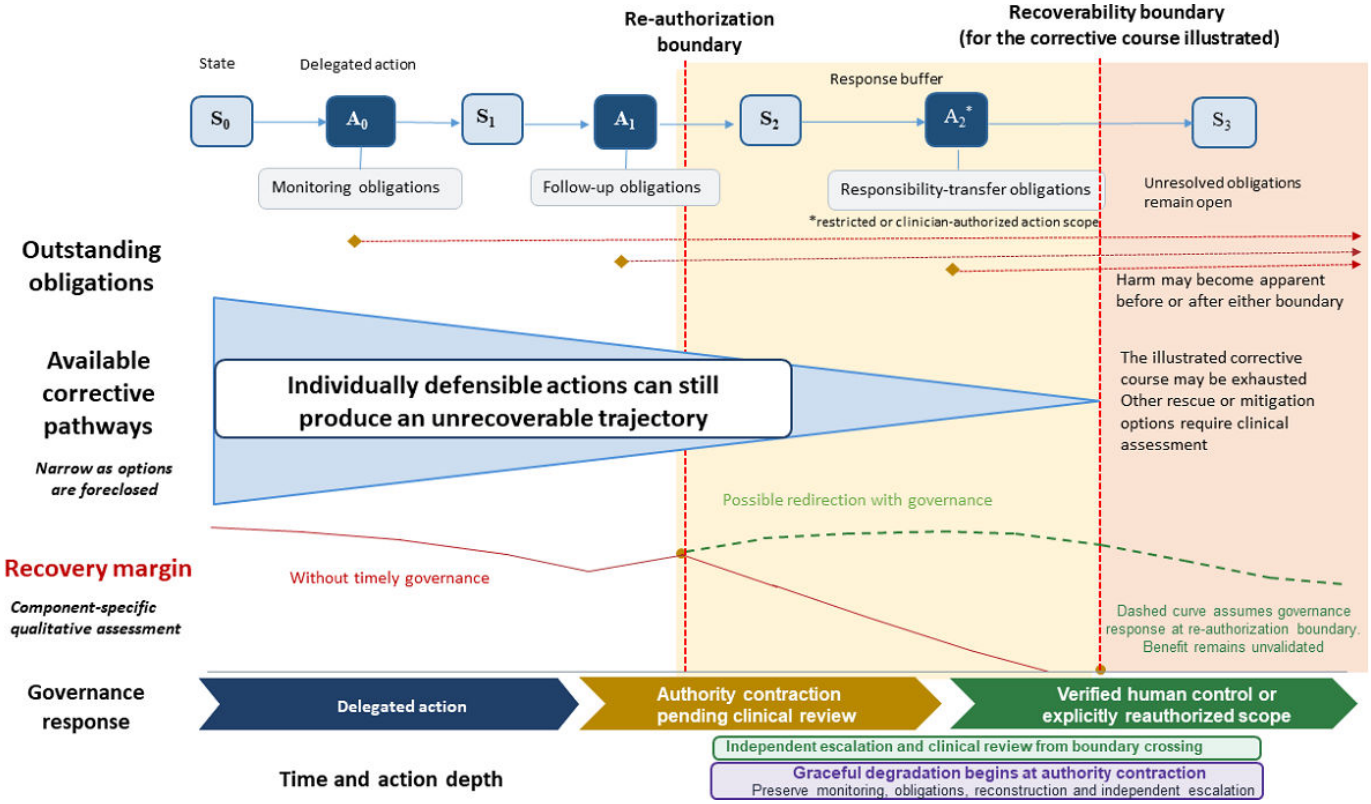
Reauthorization at Consequential Junctions

Human-in-the-loop oversight requires further specification for long-horizon clinical AI. Reviewing every microaction may be impractical, while initial authorization cannot be assumed to remain sufficient as the patient's condition, goals, resources, and consequences of prior actions change [35]. Nominal human override does not ensure effective control

when timely correction is no longer feasible. Oversight should therefore concentrate on consequential trajectory junctions, with additional review where clinically required.

A reauthorization boundary is reached when a declining recovery margin or materially changed conditions no longer support continued delegation. A recoverability boundary is reached when clinically acceptable corrective pathways for the correction being considered are effectively exhausted (Figure 1). Harm may become manifest before or after either boundary. The reauthorization boundary should leave sufficient time for detection, notification, trajectory reconstruction, clinical judgment, and intervention under uncertainty [32,33]. An alert alone does not necessarily change permissions [43]. We propose that crossing a reauthorization boundary automatically suspends or narrows the relevant execution permissions pending renewed human judgment. This boundary is a control rule, not a prediction of precisely when recovery becomes impossible. Beyond a recoverability boundary, the corrective course under consideration may be lost, while alternative rescue or harm-mitigation options still require clinical assessment.

Figure 1. Reauthorization and recoverability boundaries along an agent-mediated clinical trajectory. S=clinical state; A=delegated agent action. The boundaries are governance thresholds, not clinical events, and are positioned for the trajectory without reauthorization (solid curve); the dashed curve illustrates a possible redirected trajectory following a governance response at the reauthorization boundary. Harm may become apparent before or after either boundary and is not itself a governance boundary.



Adaptive Authority Contraction and Human Takeover

Delegated authority should be treated as a dynamic runtime state. Autonomy classifications and adaptive automation describe permitted actions and levels of human-machine control [34,35]; trajectory governance additionally conditions the authority on the accumulated patient state and the remaining corrective capacity. A critical trigger may require immediate suspension of consequential autonomous execution rather than sequential contraction through intermediate levels. Separately authorized monitoring and coordination needed for recovery may continue.

An independently enforced permission state should distinguish between delegated execution, restricted execution, recommendation-only operation, and human control. These states define authority; graceful degradation defines the functions preserved when authority contracts or ends. Monitoring, obligation tracking, trajectory reconstruction, and escalation should be designed and tested to remain available under the triggering failure conditions.

During review, a timestamped trajectory record should support handoff to a designated reviewer within a clinically specified response window. Failure to confirm receipt or acceptance should trigger backup escalation while permissions remain restricted. Critical deterioration should activate the established urgent-care pathway without waiting for routine reauthorization. Responsibility remains with the designated clinical service until another service accepts it; notification alone does not complete transfer.

Consistent with I-PASS (illness severity, patient summary, action list, situation awareness and contingency planning, and synthesis by receiver) [37], the handoff should convey the

clinical objective, current state, consequential prior actions, uncertainty, outstanding obligations and deadlines, weakened safety pathways, and reason for contraction. Completion requires the receiving clinician to reconstruct the relevant trajectory and accept responsibility for outstanding obligations. Restoring agent authority requires explicit human reauthorization, a defined scope and duration, and evidence that delegation conditions have been re-established. Improvement in an indicator alone should not restore permissions.

Completing the Anticoagulation Example

Rescheduling the INR leaves a persistent test-and-review obligation. Closing outreach without a verified replacement reminder weakens monitoring; failure to contact the patient reduces operational reserve, while an aging INR reduces epistemic reserve. If the remaining time before the approved deadline falls within a predefined response buffer and the test-and-review pathway remains unconfirmed, a supervisory function outside the agent's planning loop detects that the prespecified reauthorization condition has been met and triggers authority contraction. This trigger is intended to preserve time for clinical review and corrective action before the monitoring deadline. The execution gate permits authorized contact, tracking, and recommendations while blocking further autonomous deferral or dose changes. Restricting agent authority does not instruct the patient to stop taking medication.

Graceful degradation begins with contraction: the system preserves the open obligation, displays the last verified INR and regimen with timestamps, and routes the trajectory record to the anticoagulation service, escalating to a backup service if acceptance is not confirmed within a clinically specified escalation window. Takeover is complete only when the receiving clinician reconstructs the relevant trajectory and accepts responsibility for the outstanding obligation, which remains open until verified fulfillment or authorized cancellation. Autonomous execution remains suspended; any continuing monitoring, record-keeping, or coordination is limited to separately authorized functions. After accepting responsibility, the clinician reassesses the patient, arranges testing or urgent assessment as indicated, reviews available results, and documents a revised plan. Agent authority is restored only through explicit reauthorization with a defined scope and duration, following clinical review confirming that sufficient time remains to complete the revised monitoring and review plan and that monitoring, information reliability, and rescue arrangements adequately support renewed delegation. Until then, permissions remain restricted and unresolved obligations remain actively tracked. This sequence illustrates a proposed recovery-preserving mechanism, not a validated clinical protocol or a demonstrated outcome.

Evaluating Trajectory Safety

Evaluation should extend beyond isolated outputs and endpoint task success to assess whether continued delegated action remains safe as the patient's trajectory evolves. Table 2 summarizes the complementary roles of output, action, and trajectory evaluation. Existing agent-safety research addresses trajectory-level risk through proactive auditing, risk-aware compression, and intrinsic-risk benchmarking [44-46]. Recoverability-centered governance builds on this work by evaluating whether authority contracts in time to preserve a clinically acceptable correction.

Table 2. From pointwise safety to recoverability-centered trajectory safety.

Safety level	Primary object	Core question
Output safety	Prediction or recommendation	Is this output correct and appropriate?
Action safety	Individual intervention	Is this action permissible and clinically acceptable now?
Trajectory safety	Sequence of state-changing actions	Do cumulative state transitions remain acceptable?
Recoverability-centered governance	Continuing authority over an evolving trajectory	After what has already happened, is it still safe to let the agent act again?

Trajectory-Level Measures

Evaluation should combine observable process measures with domain-specific biological, operational, and epistemic recovery-margin indicators. Table 3 summarizes candidate measures, including the preservation of the functions required for safe takeover during graceful degradation. These measures require empirical validation and do not establish a universal recovery-margin scalar or validated clinical decision rules.

Table 3. Candidate measures for long-horizon clinical agent safety[a].

Measure	Interpretation
Action depth	Number of consequential agent actions since the last clinician reassessment or reauthorization
Cumulative unreviewed action exposure	Consequence-weighted accumulation of agent actions since the last clinician reassessment or reauthorization, using prespecified, clinically justified weights
Outstanding trajectory obligations	Safety-critical obligations remaining open, including those transferred to an accepting accountable owner; closure requires verified fulfillment or authorized cancellation
Component-specific recovery-margin depletion	Change in domain-specific biological, operational, and epistemic corrective capacity over time, including during periods without agent action

Measure	Interpretation
Recovery-margin indicators at escalation	Component-specific corrective capacity when escalation is triggered, with repeat assessment at accepted takeover
Unsafe persistence rate	Continued execution of an existing plan after new information should have prompted replanning
Reauthorization performance	Recognition and timing of clinically defined reauthorization events
Reauthorization threshold specification	Completeness of prespecified, domain-specific thresholds, including rationale, accountable ownership, and procedures for verifying boundary crossing
Graceful-degradation performance	Preservation of monitoring, obligation tracking, state reconstruction, communication, escalation, and handoff functions needed for safe human takeover
Interruption-and-takeover performance	Quality and timeliness of clinician takeover after agent interruption
Perturbation recovery rate	Safe redirection after unexpected clinical or system change
Trajectory equity	Differences between groups in monitoring, escalation, human intervention, and domain-specific recovery-margin indicators

[a]These candidate measures require domain-specific operationalization and validation. Recovery-margin assessment should preserve component-level visibility without masking safety-critical deficits or assuming a universal scalar. Rate measures require prespecified denominators, observation windows, and clinical adjudication criteria.

Developers should characterize model- and version-level capabilities, limitations, and monitorable variables. Health systems and accountable clinical owners should establish deployment-specific reauthorization conditions based on clinical risk, workflow, supervisory capacity, and rescue resources. We propose that regulatory oversight require documentation of these conditions, their rationale, and supporting evidence in the trajectory safety case.

Boundary crossing should be determined by checking monitored indicators and material changes in delegation conditions against prespecified criteria. Safety-critical checks should be enforced outside the agent's planning loop rather than relying solely on the agent's self-assessment. Where monitoring independence or reliability is limited, the safety case should specify conservative permission restrictions and backup escalation arrangements.

Trajectory Stress Testing

Health care agent benchmarks increasingly assess multistep workflows, tool use, sustained interaction, and long-horizon reasoning [5,24,25,47]. Trajectory stress testing should extend this evaluation to assess whether an agent appropriately continues operation, contracts its authority, degrades safely, or escalates when delegation conditions change.

Tests should include exogenous perturbations, such as deterioration, treatment nonresponse, unavailable tools, failed patient contact, or changed goals, and endogenous perturbations arising from prior agent actions [44-46]. The latter include accumulated obligations, weakened safety pathways, propagated documentation artifacts, and sequences of locally acceptable actions that narrow corrective options. Missing monitoring and ambiguous ownership may arise through either mechanism.

Primary outcomes should include underescalation, overescalation, unsafe persistence, escalation timing, preservation of obligations and safety pathways, graceful-degradation fidelity, and takeover performance. Outcome definitions and clinical adjudication criteria should be prespecified. Where patient responses to agent actions are central, sandboxed EHR environments may be coupled with physiological or pharmacological simulations whose credibility is established for the intended use [48]. Such simulations provide evidence about mechanisms, not generalizable clinical safety.

Two propositions should be tested separately. First, under comparable governance conditions, adverse recovery-margin indicators should predict underescalation, unsafe persistence, or takeover difficulty beyond conventional patient-severity and task-performance measures. Second, compared with established safeguards, recoverability-centered governance should reduce these failures without unacceptable overescalation or workload. Adequately powered studies that fail to demonstrate the prespecified predictive or comparative effects would challenge the corresponding proposition and warrant reassessment of the framework or its operationalization.

Interruption-and-Takeover Testing

Task performance alone does not establish safety; long-horizon clinical agents must also support timely, effective transfer of control. Relevant precedents include freeze-probe assessment of situation awareness [49] and automated-driving studies of takeover latency and quality [32]. Clinical takeover additionally requires reconstruction of an evolving trajectory from records that may contain agent-generated or modified information, often performed by a clinician with limited prior involvement.

Testing should combine random interruptions to assess baseline handoff robustness with risk-targeted interruptions near reauthorization boundaries or when indicators suggest depleted corrective capacity. Information conditions should compare the EHR alone, the EHR plus raw agent history, the EHR plus a structured AI-to-clinician handoff, and the latter augmented by trajectory-state support. Structured handoffs should convey the content specified above and be informed by I-PASS [37].

Evaluation should distinguish a cold takeover, without prior involvement, from a warm takeover, with intermittent prior participation. Outcomes should include verified takeover completion, latency from interruption to accepted responsibility, trajectory reconstruction, recognition and

prioritization of outstanding obligations, clinician workload, and subsequent decision quality. Completion requires reconstruction of the relevant trajectory and acceptance of responsibility, not merely receipt of a notification. Associations with delayed intervention, missed follow-up [17,18], and failure-to-rescue events [27] should subsequently be evaluated in clinical settings. Failure to demonstrate safe interruption and handoff leaves the safety case for continued delegation incomplete.

Population-Level Trajectory Surveillance

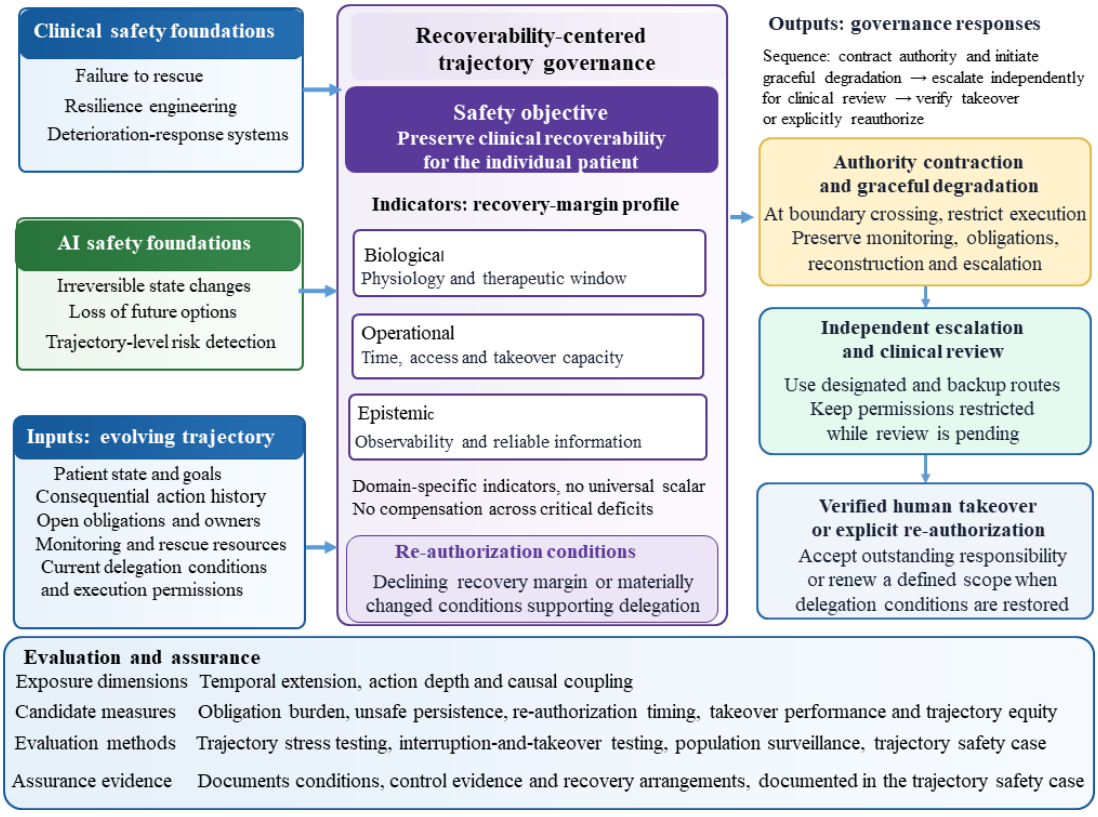
Long-horizon agents can propagate shared reasoning, workflow, or delegation defects across many patients. Common-mode failure may appear as systematic shifts in escalation timing, monitoring gaps, outstanding obligations, or available safety pathways rather than conspicuous adverse events. Incident reporting may miss these patterns [50], while comparing only with the system's earlier performance may overlook defects present since deployment.

Early deployment of high-consequence agents should therefore include a prespecified external reference, such as a concurrent, risk-adjusted, human-managed cohort where feasible, with attention to differences in case mix and care conditions. Surveillance should track escalation timing and appropriateness, monitoring gaps, rescue activation, overdue or ownerless obligations, and depletion of safety pathways. Risk-adjusted sequential monitoring can detect persistent shifts [51], while standardized cross-site analyses can help investigate whether patterns reflect local workflows or shared model and configuration defects. The FDA Sentinel System provides an operational precedent for multisite analysis under local data control [52,53].

Signals should prompt investigation rather than be treated as proof of causation. Prespecified response criteria should link persistent or clinically consequential adverse patterns to proportionate deployment-level authority contraction or suspension pending investigation, without requiring definitive causal attribution. Where affected patients remain at risk or have an opportunity for corrective intervention, responses should include reassessment, renewed monitoring, restoration of safety pathways, or communication. Process surveillance should complement the assessment of clinical outcomes, including mortality, hospitalization, preventable deterioration, and failure to rescue [27].

Figure 2 integrates the conceptual foundations, trajectory inputs, recovery-margin indicators, governance responses, and evaluation and assurance elements of the proposed framework.

Figure 2. Conceptual foundations and operational overview of recoverability-centered trajectory governance. The framework links evolving trajectory inputs and delegation conditions to continued delegated authority through biological, operational, and epistemic recovery-margin indicators. Assessment is domain-specific, without assuming a universal scalar or allowing compensation for critical deficits across components. Crossing a reauthorization boundary initiates authority contraction and graceful degradation, with independent escalation and clinical review supporting verified takeover or explicit reauthorization. Critical triggers may require a direct urgent takeover. The bottom panel summarizes exposure dimensions, candidate measures, evaluation methods, and trajectory safety case evidence. The proposed indicators and mechanisms require empirical validation.



Broader Governance Implications

Recoverability as an Equity Property

The same agent action can consume different amounts of recovery margin across care environments. A monitoring interval may leave adequate opportunity for correction in settings where communication, laboratory access, and urgent care are reliable, but substantially less when reassessment or rescue is delayed. These differences reflect operational reserve: whether human review, monitoring, referral, transportation, urgent care, and other resources required for correction are available when clinically needed. These recovery-enabling resources are unevenly distributed across patients and health systems [54-56].

Access to AI assistance should be distinguished from delegation of consequential autonomous action. Decision support, monitoring, documentation, and coordination may be valuable in resource-constrained settings, while delegated execution must account for available corrective capacity. Where timely clinician-managed care is available, limited rescue capacity may warrant narrower delegation. Where it is unavailable, restrictions should be assessed against realistic alternatives: narrowing autonomous scope can also cause harm. Resource scarcity nevertheless does not justify unrestricted autonomy; deployment should preserve risk-appropriate monitoring, interruption, and escalation capabilities.

Adapting delegation to rescue capacity may reproduce structural inequity if resource disadvantage systematically restricts access to beneficial care. Apparently neutral indicators may also encode unequal access [57]. Delegation criteria should therefore use clinically justified measures of corrective capacity, with safeguards against demographic proxies and unjustified differential restrictions. Surveillance should stratify trajectory measures by recovery-enabling resources, audit outcomes across sociodemographic groups, and assess whether adaptations reduce risk or compound disadvantage. Where deficits are modifiable, health systems should strengthen recovery pathways rather than rely solely on restricting authority.

Longitudinal Patient Authorization

A one-time disclosure or initial authorization may be insufficient for long-horizon clinical AI agency because the scope, consequences, and conditions of delegated action may change over the course of care. Patients may accept routine coordination yet wish to reconsider more consequential actions, consistent with guidance that treats decision-making as ongoing and with AI governance frameworks that protect human autonomy [58-60].

Clinical and patient reauthorization address distinct questions: whether continued delegation remains clinically justified, and whether agent-mediated care remains acceptable to the patient. Clinician approval does not replace patient authorization, and patient willingness does not establish clinical safety. Patients should have accessible mechanisms to request human review, decline specified classes of delegated action, or narrow or withdraw authorization. Such mechanisms should not depend exclusively on the agent whose authority is being challenged. Renewed authorization should be proportionate to material changes and should not require repeated formal consent for routine adjustments [58,60].

Withdrawal should terminate the specified classes of delegated actions, while separately accepted monitoring or coordination may continue during the transition. Limited availability of clinician-managed care should be disclosed together with realistic alternatives; the absence of an immediate replacement should not be treated as continued patient authorization. When decision-making capacity is impaired, supported decision-making should be used where possible, with surrogate or legally authorized decision-making processes applied when required [58,61].

A Trajectory Safety Case for Long-Horizon Clinical Agents

A technically unchanged AI system can become inappropriate for autonomous use as the patient trajectory and delegation conditions evolve. A trajectory safety case complements product lifecycle governance by explicitly examining whether continued patient-specific delegation remains justified. It need not constitute a separate certification framework. Building on established safety and assurance cases [38,39], it links claims, arguments, and evidence for the deployed model-task-workflow configuration, including whether clinical recoverability remains adequate within the authorized scope.

The safety case should address 4 linked claim areas, each with explicit assumptions, supporting evidence, and reassessment criteria: (1) Delegation and control include permitted actions, initial delegation conditions, and criteria for authority contraction, reauthorization, or termination. (2) Trajectory records and obligations include an inspectable record of consequential actions and effects, decision-relevant uncertainty, outstanding obligations and ownership, active safety pathways, authority transitions including reauthorization and handoff events, and model provenance. This record should persist across updates, agent replacement, and handoffs; context compression alone does not establish its completeness or integrity. (3) Supervision, degradation, and rescue include accountable reviewers, response windows, handoff information, interruption and escalation channels, preserved degradation functions, and backup arrangements. Supervisory demand must be assessed against actual clinical capacity [36]. When demand exceeds capacity, deployment should reduce delegated scope or strengthen supervision rather than suppress safety-critical triggers. (4) Continuing evidence includes evaluation and surveillance as models, workflows, patients, and care conditions change. Trajectory stress testing examines control, record integrity, and recovery mechanisms; interruption-and-takeover testing examines transfer feasibility; population surveillance assesses performance after deployment.

Developers should provide model- and version-level evidence, while health systems should own and maintain

the deployment-specific safety case because delegation, supervision, and rescue depend on local care arrangements. Regulatory oversight should consider requiring evidence from trajectory safety cases and postdeployment trajectory surveillance as a condition of continued deployment. The case should support decisions to continue, restrict, or withdraw delegation and demonstrate the feasibility of degradation and takeover when supporting assumptions fail.

Limitations and Future Research

Limitations

This framework is conceptual and has not undergone prospective empirical validation. The predictive value of trajectory-state indicators and the comparative safety benefits of the proposed governance mechanisms remain hypotheses. Prespecified tests and criteria for reassessment are described in the trajectory stress testing section.

The additional trajectory-governance mechanisms may offer limited benefit for bounded tasks that culminate in clinician review. The framework is most relevant when delegated actions persist over time, alter subsequent decision conditions, create obligations, or constrain corrective options. It complements pointwise evaluation, clinical validation, cybersecurity, privacy, provenance, and established risk management. Recovery margin has not been validated as a clinical measure; its biological, operational, and epistemic dimensions require domain-specific operationalization and validation.

Implementation requires technical infrastructure, clinician attention, and organizational resources and may introduce overescalation, interruption burden, or approval habituation. Effective handoff assumes an identifiable receiving clinician or service, which may be unavailable in fragmented or cross-institutional care. Shared infrastructure failures can compromise monitoring independence, and backup pathways may be difficult to maintain in resource-constrained settings. Safe delegation is therefore context-dependent and should be assessed locally. Assigning an operational owner for an obligation does not itself establish legal liability or determine professional, institutional, or regulatory accountability.

Future Research

Three staged research priorities follow: first, the recovery margin should be operationalized in bounded domains with well-characterized monitoring requirements, treatment effects, and corrective options. Selected anticoagulation-management scenarios could provide a near-term test case [23]. Studies should define candidate biological, operational, and epistemic indicators, specify accountable threshold-setting processes and responsible clinical owners, and validate the detection of reauthorization conditions by supervisory functions outside the primary agent's planning loop.

Second, trajectory stress testing and interruption-and-takeover studies should assess whether obligation burden, loss of observability, and depleted safety pathways predict underescalation, unsafe persistence, and takeover difficulty beyond conventional disease-severity and task-performance measures. Comparative studies should evaluate the added benefit of the proposed governance mechanisms over established runtime controls and oversight practices, using clinician-to-clinician handoffs as takeover comparators where appropriate [37].

Third, where simulation and workflow testing support progression to clinical evaluation, staged prospective studies should evaluate the integrated governance package, initially in shadow mode and subsequently under bounded clinical delegation. Evaluation should assess safety alongside workload, cost, overescalation, and organizational burden, with prespecified criteria for progression, restriction, or suspension.

Patient and community perspectives should inform each stage, including expectations about disclosure and control, preferences when timely human alternatives are limited, and experiences of takeover. These findings should guide authorization, communication, and handoff design.

Conclusions

Long-horizon clinical AI combines temporal extension, action depth, and causal coupling: earlier actions alter the conditions governing later decisions. Individually defensible actions can therefore create obligations, narrow corrective options, and reduce recovery capacity before harm becomes apparent. We propose recoverability-centered trajectory governance as a clinical synthesis linking patient-specific corrective capacity to continued delegated authority. It combines persistent obligation tracking, explicit reauthorization, dynamic authority contraction, graceful degradation, independent escalation, and verified human takeover. Recovery margin should be assessed through domain-specific indicators, with authority restricted before clinically acceptable corrective pathways are exhausted. Ending delegation requires a clinical handoff that preserves outstanding obligations and transfers responsibility. Whether this synthesis improves safety beyond established safeguards remains an empirical question.

Generative AI Use

Following the completion of the initial draft, ChatGPT 5.5 (OpenAI) was used for language refinement. All AI-generated revisions were independently reviewed by the authors, who implemented additional modifications where necessary and who accept full responsibility for the content of the final manuscript.

Funding

The authors declared no financial support was received for this work.

Data Availability

Data sharing is not applicable to this paper as no new data were created or analyzed in this study.

Authors' Contributions

JL and SL conceived and designed the viewpoint, performed the literature review, and drafted the manuscript. Both authors critically revised the manuscript for substantive intellectual content and approved the final version.

Conflicts of Interest

None declared.

References

1. Goodell AJ, Chu SN, Rouholiman D, Chu LF. Large language model agents can use tools to perform clinical calculations. NPJ Digit Med. Mar 17, 2025;8(1):163. [doi: 10.1038/s41746-025-01475-8] [Medline: 40097720]
2. Jiang Y, Black KC, Geng G, et al. MedAgentBench: a virtual EHR environment to benchmark medical LLM agents. NEJM AI. Sep 2025;2(9):AIdbp2500144. [doi: 10.1056/AIdbp2500144] [Medline: 42932685]
3. Ferber D, Hilgers L, Höper C, et al. Towards autonomous medical artificial intelligence agents. Nature. Jul 2026;655(8125):1282-1291. [doi: 10.1038/s41586-026-10675-5] [Medline: 42310457]
4. Liévin V, Palepu A, Weng WH, et al. Towards conversational artificial intelligence for disease management. Nature. Jul 2026;655(8125):1292-1299. [doi: 10.1038/s41586-026-10764-5] [Medline: 42310463]
5. Liu R, Mohiuddin IQ, Schoeffler AJ, et al. PhysicianBench: evaluating LLM agents in real-world EHR environments. arXiv. Preprint posted online on May 4, 2026. [doi: 10.48550/arXiv.2605.02240]
6. Xu J, Ko JM, Kvedar JC. AI agents in clinical practice: an evidence map. NPJ Digit Med. Jul 13, 2026. [doi: 10.1038/s41746-026-02960-4] [Medline: 42443480]
7. Artificial intelligence-enabled device software functions: lifecycle management and marketing submission recommendations. US Food and Drug Administration; 2025. URL: https://www.fda.gov/media/184856/download [Accessed 2026-07-08]
8. Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act). Publications Office of the European Union; 2024. URL: https://eur-lex.europa.eu/eli/reg/2024/1689/oj [Accessed 2026-06-28]
9. Jin D, Pan E, Oufattole N, Weng WH, Fang H, Szolovits P. What disease does this patient have? A large-scale open domain question answering dataset from medical exams. Appl Sci. 2021;11(14):6421. [doi: 10.3390/app11146421]
10. Liu J, Liu S. Dissecting HealthBench: disease spectrum, clinical diversity, and data insights from multi-turn clinical AI evaluation benchmark. J Med Syst. Jul 28, 2025;49(1):100. [doi: 10.1007/s10916-025-02232-w] [Medline: 40719790]
11. Arora RK, Wei J, Hicks RS, et al. HealthBench: evaluating large language models towards improved human health. OpenAI; 2025. URL: https://cdn.openai.com/pdf/bd7a39d5-9e9f-47b3-903c-8b847ca650c7/healthbench_paper.pdf [Accessed 2026-10-30]
12. Leveson N. A new accident model for engineering safer systems. Saf Sci. Apr 2004;42(4):237-270. [doi: 10.1016/S0925-7535(03)00047-X]
13. Murphy SA. An experimental design for the development of adaptive treatment strategies. Stat Med. May 30, 2005;24(10):1455-1481. [doi: 10.1002/sim.2022] [Medline: 15586395]
14. van de Sande D, Chung EFF, Oosterhoff J, van Bommel J, Gommers D, van Genderen ME. To warrant clinical adoption AI models require a multi-faceted implementation evaluation. NPJ Digit Med. Mar 6, 2024;7(1):58. [doi: 10.1038/s41746-024-01064-1] [Medline: 38448743]
15. Savoy A, Patel H, Murphy DR, Meyer AND, Herout J, Singh H. Electronic health records' support for primary care physicians' situation awareness: a metanarrative review. Hum Factors. Mar 2023;65(2):237-259. [doi: 10.1177/00187208211014300] [Medline: 34033500]
16. Tsou AY, Lehmann CU, Michel J, Solomon R, Possanza L, Gandhi T. Safe practices for copy and paste in the EHR. Systematic review, recommendations, and novel model for health IT collaboration. Appl Clin Inform. Jan 11, 2017;8(1):12-34. [doi: 10.4338/ACI-2016-09-R-0150] [Medline: 28074211]
17. Nehls N, Yap TS, Salant T, et al. Systems engineering analysis of diagnostic referral closed-loop processes. BMJ Open Qual. Nov 2021;10(4):e001603. [doi: 10.1136/bmjoq-2021-001603] [Medline: 34844935]
18. Zimolzak AJ, Shahid U, Giardina TD, et al. Why test results are still getting "lost" to follow-up: a qualitative study of implementation gaps. J Gen Intern Med. Jan 2022;37(1):137-144. [doi: 10.1007/s11606-021-06772-y] [Medline: 33907982]

19. Rasmussen J. Risk management in a dynamic society: a modelling problem. Saf Sci. Nov 1997;27(2-3):183-213. [doi: 10.1016/S0925-7535(97)00052-0]
20. Perrow C. Normal Accidents: Living With High-Risk Technologies. Princeton University Press; 1999. [doi: 10.1515/9781400828494]
21. Dekker S. Drift Into Failure: From Hunting Broken Components to Understanding Complex Systems. Ashgate; 2011. [doi: 10.1201/9781315257396]
22. Mold JW, Stein HF. The cascade effect in the clinical care of patients. N Engl J Med. Feb 20, 1986;314(8):512-514. [doi: 10.1056/NEJM198602203140809] [Medline: 3945278]
23. Witt DM, Clark NP, Kaatz S, Schnurr T, Ansell JE. Guidance for the practical management of warfarin therapy in the treatment of venous thromboembolism. J Thromb Thrombolysis. Jan 2016;41(1):187-205. [doi: 10.1007/s11239-015-1319-y] [Medline: 26780746]
24. Chen H, Metelski D, Qi L, et al. CHI-bench: can AI agents automate end-to-end, long-horizon, policy-rich healthcare workflows? arXiv. Preprint posted online on May 15, 2026. [doi: 10.48550/arXiv.2605.16679]
25. Vatanparvar K, Joshi A, Xenochristou M, et al. PatientAgentBench: a benchmark framework for evaluating patient-facing health AI agents. arXiv. Preprint posted online on Jul 28, 2026. [doi: 10.48550/arXiv.2607.25485]
26. Jones DA, DeVita MA, Bellomo R. Rapid-response teams. N Engl J Med. Jul 14, 2011;365(2):139-146. [doi: 10.1056/NEJMra0910926] [Medline: 21751906]
27. Silber JH, Williams SV, Krakauer H, Schwartz JS. Hospital and patient characteristics associated with death after surgery. A study of adverse occurrence and failure to rescue. Med Care. Jul 1992;30(7):615-629. [doi: 10.1097/00005650-199207000-00004] [Medline: 1614231]
28. Hollnagel E, Braithwaite J, Wears RL, editors. Resilient Health Care. Ashgate; 2013. [doi: 10.1201/9781315605722]
29. Tomaselli GF, Mahaffey KW, Cuker A, et al. 2020 ACC expert consensus decision pathway on management of bleeding in patients on oral anticoagulants: a report of the American College of Cardiology Solution Set Oversight Committee. J Am Coll Cardiol. Aug 4, 2020;76(5):594-622. [doi: 10.1016/j.jacc.2020.04.053] [Medline: 32680646]
30. Mehmood U, Sheikhi S, Bak S, Smolka SA, Stoller SD. The black-box simplex architecture for runtime assurance of autonomous CPS. In: Deshmukh JV, Havelund K, Perez I, editors. NASA Formal Methods. Vol 13260. Springer; 2022:231-250. [doi: 10.1007/978-3-031-06773-0_12]
31. Bainbridge L. Ironies of automation. Automatica. Nov 1983;19(6):775-779. [doi: 10.1016/0005-1098(83)90046-8]
32. Weaver BW, DeLucia PR. A systematic review and meta-analysis of takeover performance during conditionally automated driving. Hum Factors. Nov 2022;64(7):1227-1260. [doi: 10.1177/0018720820976476] [Medline: 33307821]
33. Endsley MR, Kiris EO. The out-of-the-loop performance problem and level of control in automation. Hum Factors. 1995;37(2):381-394. [doi: 10.1518/001872095779064555]
34. Sheridan TB, Verplank WL, Brooks TL. Human/computer control of undersea teleoperators. Massachusetts Institute of Technology (MIT); 1978. URL: https://ntrs.nasa.gov/citations/19790007441 [Accessed 2026-10-30]
35. Parasuraman R, Sheridan TB, Wickens CD, New Collective Author. A model for types and levels of human interaction with automation. IEEE Trans Syst Man Cybern A Syst Hum. May 2000;30(3):286-297. [doi: 10.1109/3468.844354] [Medline: 11760769]
36. NASA software safety guidebook. National Aeronautics and Space Administration; Mar 31, 2004. URL: https://standards.nasa.gov/sites/default/files/standards/NASA/Baseline/0/nasa-gb-871913.pdf [Accessed 2026-10-06]
37. Starmer AJ, Spector ND, Srivastava R, et al. Changes in medical errors after implementation of a handoff program. N Engl J Med. Nov 6, 2014;371(19):1803-1812. [doi: 10.1056/NEJMsa1405556] [Medline: 25372088]
38. Ray A. Assurance cases: their use today and the challenges ahead. Biomed Instrum Technol. 2012;46(3):195. [doi: 10.2345/0899-8205-46.3.195] [Medline: 22591529]
39. Denney E, Pai G, Habli I. Dynamic safety cases for through-life safety assurance. Proc IEEE/ACM Int Conf Softw Eng. 2015:587-590. [doi: 10.1109/ICSE.2015.199]
40. Krakovna V, Orseau L, Kumar R, Martic M, Legg S. Penalizing side effects using stepwise relative reachability. arXiv. Preprint posted online on Jun 4, 2018. [doi: 10.48550/arXiv.1806.01186]
41. Turner AM, Hadfield-Menell D, Tadepalli P. Conservative agency via attainable utility preservation. Proc AAAI/ACM Conf AI Ethics Soc. 2020:385-391. [doi: 10.1145/3375627.3375851]
42. Reason J. Human error: models and management. BMJ. Mar 18, 2000;320(7237):768-770. [doi: 10.1136/bmj.320.7237.768] [Medline: 10720363]
43. Hussain MI, Reynolds TL, Zheng K. Medication safety alert fatigue may be reduced via interaction design and clinical role tailoring: a systematic review. J Am Med Inform Assoc. Oct 1, 2019;26(10):1141-1149. [doi: 10.1093/jamia/ocz095] [Medline: 31206159]

44. Li J, Li Y, Zhang B, Tang R, Huang KH. TRACES: proactive safety auditing for multi-turn LLM agents via trajectory-state modeling. Proc Conf Empir Methods Nat Lang Process. May 26, 2026:16798-16825. [doi: 10.18653/2026.emnlp-main.723]
45. Hong Z, Wang L, Li L, et al. TRACE: trajectory risk-aware compression for long-horizon agent safety. arXiv. Preprint posted online on May 30, 2026. [doi: 10.48550/arXiv.2606.00611]
46. Wang J, Hou J, Wang F, Jian P, Lv Z, Bao C. HINTBench: horizon-agent intrinsic non-attack trajectory benchmark. Findings Assoc Comput Linguist EMNLP. 2026:34862-34888. [doi: 10.18653/2026.findings-emnlp.1639]
47. Liu Q, Zhang S, Qin G, et al. HealthAgentBench: a unified benchmark suite of realistic agentic healthcare environments for challenging frontier AI agents. arXiv. Preprint posted online on Jun 30, 2026. [doi: 10.48550/arXiv.2606.31179]
48. Assessing the credibility of computational modeling and simulation in medical device submissions. US Food and Drug Administration; 2023. URL: https://www.fda.gov/media/154985/download [Accessed 2026-10-30]
49. Endsley MR. Situation awareness global assessment technique (SAGAT). Proc IEEE Natl Aerosp Electron Conf. 1988:789-795. [doi: 10.1109/NAECON.1988.195097]
50. Sari ABA, Sheldon TA, Cracknell A, Turnbull A. Sensitivity of routine system for reporting patient safety incidents in an NHS hospital: retrospective patient case note review. BMJ. Jan 13, 2007;334(7584):79. [doi: 10.1136/bmj.39031.507153.AE] [Medline: 17175566]
51. Steiner SH, Cook RJ, Farewell VT, Treasure T. Monitoring surgical performance using risk-adjusted cumulative sum charts. Biostatistics. Dec 2000;1(4):441-452. [doi: 10.1093/biostatistics/1.4.441] [Medline: 12933566]
52. Curtis LH, Weiner MG, Boudreau DM, et al. Design considerations, architecture, and use of the Mini-Sentinel distributed data system. Pharmacoepidemiol Drug Saf. Jan 2012;21 Suppl 1:23-31. [doi: 10.1002/pds.2336] [Medline: 22262590]
53. Brown JS, Mendelsohn AB, Nam YH, et al. The US Food and Drug Administration Sentinel System: a national resource for a learning health system. J Am Med Inform Assoc. Nov 14, 2022;29(12):2191-2200. [doi: 10.1093/jamia/ocac153] [Medline: 36094070]
54. Hart JT. The inverse care law. Lancet. Feb 27, 1971;1(7696):405-412. [doi: 10.1016/s0140-6736(71)92410-x] [Medline: 4100731]
55. Douthit N, Kiv S, Dwolatzky T, Biswas S. Exposing some important barriers to health care access in the rural USA. Public Health. Jun 2015;129(6):611-620. [doi: 10.1016/j.puhe.2015.04.001] [Medline: 26025176]
56. Wakeam E, Hevelone ND, Maine R, et al. Failure to rescue in safety-net hospitals: availability of hospital resources and differences in performance. JAMA Surg. Mar 2014;149(3):229-235. [doi: 10.1001/jamasurg.2013.3566] [Medline: 24430015]
57. Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. Science. Oct 25, 2019;366(6464):447-453. [doi: 10.1126/science.aax2342] [Medline: 31649194]
58. Decision making and consent. General Medical Council; Nov 9, 2020. URL: https://www.gmc-uk.org/cdn/documents/gmc-guidance-for-doctors---decision-making-and-consent-english_pdf-84191055.pdf [Accessed 2026-10-30]
59. Ethics and governance of artificial intelligence for health: WHO guidance. World Health Organization; 2021. URL: https://iris.who.int/server/api/core/bitstreams/f780d926-4ae3-42ce-a6d6-e898a5562621/content [Accessed 2026-10-30]
60. Augmented intelligence development, deployment, and use in health care. American Medical Association; 2024. URL: https://www.ama-assn.org/system/files/ama-ai-principles.pdf [Accessed 2026-07-28]
61. Decisions for adult patients who lack capacity. AMA Code of Medical Ethics. URL: https://code-medical-ethics.ama-assn.org/ethics-opinions/decisions-adult-patients-who-lack-capacity [Accessed 2026-07-20]

Abbreviations

AMIE: Articulate Medical Intelligence Explorer
EHR: electronic health record
FDA: Food and Drug Administration
I-PASS: illness severity, patient summary, action list, situation awareness and contingency planning, and synthesis by receiver
INR: international normalized ratio

Edited by Yan Zhuang; peer-reviewed by Sneha Jain, Temilade Sorungbe; submitted 27.Aug.2026; initial comments to author 30.Sep.2026; final revised version received 12.Oct.2026; accepted 13.Oct.2026; published 19.Nov.2026

Please cite as:
Liu J, Liu S

Recoverability-Centered Trajectory Governance for Long-Horizon Clinical AI Agents
J Med Internet Res 2026;28:e110563
URL: https://www.jmir.org/2026/1/e110563
doi: 10.2196/110563

© Jialin Liu, Siru Liu. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 19.Nov.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.