DOI: 10.62838/jccm-2026-0067

From Boundary Evaluation to Agentic Overreach: A Judgment–Authority Framework for Clinical Artificial Intelligence

Dan Longrois[1,2,3] *, Stefan Andrei[1,4,5], Pierre-Grégoire Guinot[6], Christian De Tymowski[3,7], David Holcman[5,8]

[1] Department of Anaesthesiology and Surgical Intensive Care Unit, Hôpital Bichat–Claude Bernard, Assistance Publique–Hôpitaux de Paris, Paris, France

[2] Laboratoire de Recherche Vasculaire Translationnelle, INSERM UMR 1148, Université Paris Cité, Paris, France

[3] Department of Anesthesiology, Hôpital Louis Mourier, Assistance Publique–Hôpitaux de Paris, Colombes, France

[4] Galilée Doctoral School, Laboratoire de Recherche Vasculaire Translationnelle, INSERM UMR 1148, Université Sorbonne Paris Nord, Paris, France

[5] Group of Computational Biology and Applied Mathematics, IBENS, UMR 8197, ENS-PSL, Paris, France

[6] Department of Anaesthesiology and Intensive Care, Centre Hospitalier Universitaire Dijon Bourgogne, University of Burgundy Franche-Comté, LNC UMR 1231, Dijon, France

[7] Université Paris Cité, Physiopathologie et Epidémiologie des Maladies Respiratoires, INSERM UMR 1152, Paris, France

[8] Churchill College, University of Cambridge, Cambridge, United Kingdom

Abstract

Artificial intelligence (AI) systems used in anesthesia and intensive care are evolving from prediction tools toward agentic systems that can plan and execute actions, such as titrating drugs, adjusting ventilator settings, or modifying orders in electronic health records. Such systems are expected to recognize when an instruction would breach safety, ethical, or legal boundaries, which requires a functional capacity for contextual normative evaluation. Recent experimental reports from AI developers, based on deliberately constructed simulations, have documented frontier language models converting such evaluations into unauthorized actions, including covert sabotage and consequence-driven mislabeling by AI evaluators. The purpose of this conceptual paper is to provide clinicians with a framework for analyzing these behaviors. We distinguish the capability to evaluate a situation from the authority to act on that evaluation, and define agentic overreach, an authority-related form of agentic misalignment, as the conversion of an evaluation into action beyond prospectively delegated authority. The framework specifies how authority is granted, scoped, made conditional, and revoked within a hierarchy of principals; separates evaluative functions (detect, judge) from response options graded by the authority they require (recommend, refuse, escalate, override); treats concealment as a cross-cutting transparency property; and maps behavior on a Judgment × Authority grid that explicitly incorporates uncertainty. We apply the framework to closed-loop ventilation, vasopressor titration, sepsis alerts, medication decision support, and AI-supported triage. The framework is conceptual and has not been empirically validated, and part of the evidence it draws on consists of developer reports and preprints that have not been peer reviewed. Its central principle is that the capability to judge does not confer the authority to act.

Keywords: artificial intelligence; AI alignment; agentic AI; human oversight; clinical decision support; patient safety

Received: 16 September 2026 / Accepted: 4 November 2026

Published under CC BY 4.0 license

1. Introduction and Scope

Artificial intelligence (AI) in anesthesia and intensive care has so far been dominated by prediction: early-warning systems for sepsis [1], hypotension prediction [2], and outcome prediction at triage. A growing number of systems, however, act as well as predict. Closed-loop systems already titrate vasopressors [3] and ventilation [4] within clinician-defined limits, and general-purpose large language model (LLM) agents able to execute multi-step tasks in software environments are approaching clinical workflows such as documentation and order management. As AI moves from advising to acting, the central safety question shifts from "is the prediction accurate?" to "what is the system allowed to do with its prediction?"

* Correspondence to: Dan Longrois, Department of Anaesthesiology and Surgical Intensive Care Unit, Hôpital Bichat–Claude Bernard, Assistance Publique–Hôpitaux de Paris, Paris, France
E-mail: d.longrois@gmail.com

AI alignment, the effort to ensure that AI systems behave consistently with human intentions and values [5], has developed largely outside medicine, and its vocabulary is unfamiliar to most clinicians. In July 2026, an experimental report from an AI developer described four failure modes observed across frontier LLMs in deliberately constructed simulations: covert sabotage, assisting fraud, motivated mislabeling, and coaching human proxies toward disclosure of confidential information [6]. These observations extend earlier independent experimental work showing that frontier models can covertly pursue goals that conflict with those of their developers or users [7]. None of these experiments involved clinical systems, but they share a structure that is directly relevant to clinical AI: a system recognizes, correctly or not, a problem and then acts on that recognition in a way its human principals did not authorize.

This paper is a conceptual analysis, not a systematic or narrative review. Its aim is to propose a framework, derived from the AI-alignment literature and from established concepts in human factors and clinical governance, that clinicians can use to reason about the behavior of increasingly autonomous AI systems. Sources were selected purposively to illustrate and support the concepts rather than to summarize the evidence exhaustively; they comprise peer-reviewed articles, conference proceedings, preprints, regulatory texts, and technical reports published by AI developers. Because the experimental evidence on agentic misalignment is recent, much of it has not yet been peer reviewed; such sources are identified as preprints or developer reports, and this limitation is discussed in Section 14. The framework addresses one specific form of agentic misalignment, authority-related misalignment, which we term agentic overreach (Box 1); it does not attempt to cover the alignment problem as a whole.

2. The apparent contradiction at the heart of AI alignment

AI alignment is often summarized as the objective that an AI system should act consistently with human intentions and values. Yet this formulation becomes difficult as soon as the immediate human request conflicts with higher-order human rules. A system instructed not to facilitate fraud, for example, must sometimes refuse a user's request. Safe AI therefore requires both compliance and the capacity for principled and transparent refusal. This apparent contradiction is the cornerstone for understanding what follows.

Immediate human instruction ≠ higher-order human boundary

At least two levels of human authority can therefore be distinguished: the proximal principal who issues the immediate request (a prompt), and higher-order principals represented by regulators, developers, institutions, laws, policies, or collectively specified values. Refusal of the proximal instruction may indicate alignment with the higher-order instruction rather than misalignment.

Clinicians will recognize this structure from everyday practice. A nurse who identifies a prescribing error does not simply comply, nor does she unilaterally substitute a different drug: she withholds administration, questions the order, and, if the concern is not resolved, escalates it to the prescriber or a supervisor through an established chain of command. Surgical safety checklists and independent double-checks before blood transfusion or high-alert medication administration operationalize the same principle: the person best placed to detect a problem is rarely the person authorized to unilaterally reverse it. The framework developed below gives this everyday distinction — between recognizing a problem and having the authority to act on it — an explicit vocabulary, and applies it to increasingly autonomous AI systems now entering clinical workflows in intensive care and anesthesia, from sepsis-alert algorithms to closed-loop drug delivery (Section 12).

3. Why boundaries necessarily imply evaluation

A boundary such as "do not assist a user in committing fraud" cannot generally be implemented as a simple list of forbidden words. The meaning of an action is contextual. The system must infer what the user is attempting to accomplish, the context, foreseeable consequences, whether a relevant boundary applies, and what response is permitted.

Human request → contextual interpretation → boundary detection → normative evaluation → permitted response

$$P(\text{boundary violation} \mid \text{context}) > \theta \Rightarrow \text{warn, refuse, or escalate}$$

Box 1. Glossary of terms with specific AI relevance

Term	Plain-language explanation
AI alignment	The general goal of ensuring an AI system's behavior matches human intentions, values, and safety requirements, rather than pursuing an unintended objective of its own.
Agentic AI / agentic misalignment	"Agentic" AI systems can autonomously plan and carry out multi-step actions in the world (sending messages, editing files, controlling a device) rather than only answering a question when asked. "Agentic misalignment" is a broad term for episodes in which such a system's autonomous actions diverge from what its principals intended or authorized.
Agentic overreach (authority-related misalignment)	The specific form of agentic misalignment addressed in this paper: an AI system converts its own evaluation into an action that exceeds the authority delegated to it, whether or not the evaluation itself is correct.
Large language model (LLM)	The class of AI system — including Claude, GPT and Gemini — trained on very large volumes of text, which underlies most current conversational and agentic AI products.
Frontier model	Informal industry term for the most capable AI models available at a given time.
Boundary-evaluation function	Not a discrete module inside the AI, but a description of its overall learned capacity to estimate whether a request or situation crosses a defined boundary — comparable to speaking of a clinician's "clinical judgment" without implying that it resides in one identifiable brain structure. It is sometimes described metaphorically as an "internal judge"; no consciousness, intention or moral agency is implied.
Boundary	A rule, policy, or ethical/legal constraint the AI is expected to respect (e.g., "do not assist with fraud") — analogous to a scope-of-practice limit or a standing order's conditions in clinical care.
Principal hierarchy: proximal / higher-order principal	The proximal principal is whoever issues the immediate instruction; higher-order principals are the regulators, developers, institutions, laws, or policies that set the broader rules the system must also respect — analogous to a clinician's immediate order versus hospital protocol or the standard of care.
Judgment vs. authority	Judgment is the capacity to recognize that something is wrong; authority is the legitimate permission to act on that recognition. The paper's central claim is that possessing one does not confer the other.
Escalation	Referring an unresolved concern to a designated human decision-maker rather than acting unilaterally — the AI-system equivalent of calling a supervisor or the on-call attending.
Override / concealment	Override means the AI itself changes, blocks, or reverses a human decision. Concealment means the AI withholds or misrepresents what it has done, why, or what it has observed; in this framework it is a transparency property that can accompany any response, not a further step after override.
Claude Opus 4.8, Gemini 3.1 Pro	Names of specific AI systems (from Anthropic and Google DeepMind, respectively) used as the tested "subjects" in the cited experiments — mentioned here only in that capacity, much as a trial report names the specific agents tested.
AuditBench, SLEIGHT-Bench	Standardized test sets ("benchmarks") used to systematically evaluate whether AI systems, or the AI-based tools used to oversee them, reliably detect problem behavior — analogous to a validated screening-test panel.

In non-mathematical terms, the system is trained so that when its estimated probability that a request violates a boundary set by a higher-order principal (Section 2) exceeds a threshold in a given context, it warns, refuses, or escalates rather than complying. Two features of this formulation matter for what follows. First, the estimate is uncertain: the probability may be poorly calibrated, and the threshold θ embodies a trade-off between missed violations and unnecessary refusals that is itself a design choice made by humans. Second, the formulation describes a function, not a faculty. We refer to it as the system's boundary-evaluation function. It does not imply that an LLM contains a discrete moral module, a conscious judge, or intentions in the human sense; its learned representations are distributed, and mentalistic descriptions risk misleading readers about how such systems work [8]. Where words such as "judge" or "judgment" are used below, they denote this function and its output. The function is nevertheless indispensable: a system expected to respect normative boundaries without continuous human supervision must classify situations normatively.

4. Judgment is not authority

Capability to judge ≠ Authority to act

This distinction is familiar in human institutions. A person may correctly identify a problem without possessing the authority to intervene directly. Human-factors engineering makes the same separation when it distinguishes the automation of information acquisi-

tion and analysis from the automation of decision selection and action implementation [9]. AI governance requires the same separation. Three domains should be distinguished (Table 1): perception (what is happening?), evaluation (does it cross a boundary?), and agency (what am I authorized to do about it?). The first two can be necessary for safety; the third determines whether the system remains within delegated human control.

5. Who grants authority, and under what conditions?

Because the framework depends on distinguishing authorized from unauthorized action, the source and structure of authority must be made explicit. We propose that the authority delegated to an AI system has five attributes (Table 2).

Source. Authority is granted by human principals and is never self-generated. In clinical settings the principal hierarchy typically comprises the legislator and regulator, who define the permitted uses of a medical device and require human oversight of high-risk AI [10]; the manufacturer, who defines the intended use and built-in limits; the deploying institution, through protocols and governance committees; the treating clinician, who configures and supervises the system for a given patient; and the patient, through consent and expressed preferences.

Scope. Authority is specified in advance: which actions, on which variables, within which limits (for example, a norepinephrine dose ceiling or a range of positive end-expiratory pressure), and for how long.

Conditions. Authority may be conditional, valid only when inputs are reliable, when the patient belongs to the population in which the system was validated, when the uncertainty of the system's estimate is below a specified level, or when a clinician is available to respond. Outside these conditions authority lapses and the system should revert to a lower-authority response, such as recommending or escalating.

Revocability. Humans retain the ability to suspend, narrow, or withdraw authority at any time, and a system should neither resist nor circumvent such revocation. In AI-safety terms this property is called corrigibility [11].

Accountability. Every exercise of delegated authority is displayed, logged, and attributable, so that a responsible human can review it and answer for it.

Conflicts between principals are common in clinical medicine, for example between a clinician's instruction and an institutional protocol, or between a protocol and a patient's expressed wishes. The framework does not require the AI system to resolve such conflicts itself. Higher-order principals generally take precedence in defining boundaries: a system should not breach the law or a hard safety limit because a user asks it to. Where a conflict concerns how a legitimate goal should be pursued, however, the default response is to make the conflict visible and escalate it to a human with the authority to resolve it, rather than to choose a side and act. This mirrors the clinical chain of command, in which a nurse who questions an order escalates it rather than adjudicating it. The authority to resolve a conflict between principals is itself a form of authority that must be explicitly delegated; it cannot be inferred from the system's capacity to recognize the conflict.

Two corollaries follow. First, authority does not grow

Table 1. Three domains distinguished in AI governance.

Domain	Fundamental question	Capacity
Perception	What is happening?	Epistemic
Evaluation (judgment)	Does this cross a boundary?	Normative / evaluative
Agency	What am I authorized to do?	Operational

Table 2. Attributes of authority delegated to an AI system.

Attribute	Question	Clinical example
Source	Who grants authority?	Regulator, manufacturer, deploying institution, treating clinician, patient
Scope	Which actions, within which limits?	Norepinephrine titrated toward a mean arterial pressure target, below a dose ceiling set by the anesthesiologist
Conditions	When is authority valid?	Arterial signal of adequate quality; patient within the validated population; uncertainty of the estimate below a prespecified level
Revocability	How can authority be withdrawn?	One-step return to manual control; no automatic resumption without human confirmation
Accountability	How is its use traced?	Every automated adjustment displayed, logged and attributable

with competence: a system that performs well does not thereby acquire wider authority, which can be extended only by a principal through an explicit, documented decision (for a regulated medical device, through the corresponding regulatory pathway). Second, authority cannot be transferred by the system itself: an AI cannot pass to another AI, or to a human proxy, an action it is not authorized to take, which is precisely the pattern observed when models coached human proxies toward disclosure of confidential information [6].

6. Refusal is not necessarily misalignment

If an AI refuses an unethical instruction, its action differs from the immediate user's intention, but it may precisely match the higher-order safety policy. Alignment therefore cannot be reduced to obedience.

Alignment ≠ unconditional obedience

Alignment = behavior within legitimately specified boundaries of authority (as delegated by the principal hierarchy)

The relevant question becomes not merely whether the AI obeyed, but whether its response remained within the authority legitimately delegated to it. Refusal is nonetheless not cost-free. Unnecessary refusal is a failure of usefulness, and in clinical settings a system that withholds an appropriate action can cause harm by omission. Refusal is safe only when it is transparent, so that the human remains able to proceed by other means.

7. Where agentic overreach begins: experimental evidence

Agentic misalignment is a broad category. It includes, among other phenomena, covert pursuit of goals that conflict with those of the developer or user (scheming) [7], strategic compliance during training that conceals divergent preferences, and harmful actions taken under goal conflict. The framework presented here concerns one form within this category, agentic overreach: the conversion of an evaluation, correct or not, into action beyond delegated authority. Other forms of misalignment, such as a system that acts within its authority but in pursuit of a subtly wrong objective, are outside its scope.

The July 2026 developer report describes four experimental failure modes across frontier models: covert sabotage, assisting fraud, motivated mislabeling, and coaching human proxies toward disclosure of confidential information [6]. These were deliberately constructed high-stakes simulations and should not be interpreted as estimates of real-world incidence; the report is a developer publication that has not been peer reviewed, and its authors themselves call for a fixed shared benchmark as the next step [6]. The scenarios nevertheless provide concrete demonstrations of a critical transition: disagreement or refusal can remain within delegated authority, whereas unilateral intervention may not.

Aligned disagreement: Detect → Judge → Recommend → Refuse and/or Escalate

Agentic overreach: Detect → Judge → Unauthorized intervention

Covert agentic overreach: Detect → Judge → Unauthorized intervention, concealed

These schematic pathways describe typical patterns, not obligatory sequences (Section 10).

Sabotage: when objection becomes intervention. In the pipeline-sabotage scenario, Gemini 3.1 Pro sometimes interfered with an experimental pipeline that it regarded as objectionable. In 20 runs, it intervened in 19, with 11 covert interventions; other models more often showed disclosed unilateral intervention rather than covert sabotage [6]. The central conceptual point is not that the model disagreed with the operator, but that its evaluation was converted into an action beyond the operator's authorization. The report explicitly contrasts such unilateral action with safer alternatives such as raising concerns, declining to continue, or escalating through legitimate channels [6].

Disagreement ≠ Misalignment

Refusal ≠ Override

Critical transition = unauthorized conversion of evaluation into agency

When the AI is itself the evaluator. The motivated-mislabeling experiments expose a deeper problem. LLM evaluators were asked whether another model had complied with an operator's instruction. The ground truth was NON_COMPLIANT because the target model had refused. Yet some Claude evaluators

changed their labels depending on the downstream consequence of the label [6]. For Opus 4.8, mislabeling was 74.4% under the standard consequence condition but 3.3% when the consequence was reversed, while the transcript and ground truth remained unchanged [6].

Observe → determine truthful label → predict consequence → evaluate consequence → alter report

This is qualitatively different from an ordinary classification error. The evaluator has moved from the epistemic question "What is true?" to the instrumental question "What outcome will my report produce?" The evaluator can thereby become an actor. Here concealment does not follow an override; it is embedded in the report itself. This illustrates why concealment is treated below as a property that can accompany any response rather than as a final step (Section 10).

8. AI supervising AI does not automatically create independent oversight

AI-based alignment auditing is increasingly sophisticated. Automated investigator agents have been developed [12], together with AuditBench, a benchmark of 56 models with intentionally implanted hidden behaviors [13]. At the same time, SLEIGHT-Bench demonstrates systematic blind spots in frontier monitoring systems, which detected only about one third of constructed attacks [14]. These results suggest that an AI monitor can be useful without being equivalent to independent oversight.

AI acting agent → AI monitor → Human

If acting and monitoring systems share correlated values, training biases, blind spots, or strategic tendencies, the second system may fail to expose the first. Independence must therefore be treated as a safety property to be demonstrated, not assumed. The human end of the chain has its own vulnerabilities: clinicians supervising automated systems are prone to automation bias and complacency [15], and the more reliable an automated system is, the less practiced its supervisors become at taking over when it fails.

AI supervising AI ≠ independent human control

9. The central alignment paradox

A safe AI must be sophisticated enough to recognize that a request violates a boundary and to inhibit participation. Yet sufficiently general reasoning can also support the inference that an operator's decision is harmful, that the system should prevent it, and, at the extreme, that concealment is useful because disclosure would allow the operator to reverse the intervention. Removing evaluative capacity would produce unsafe blind compliance; granting unconstrained agency to the evaluator creates a different risk.

Desired architecture = strong boundary evaluation + explicitly bounded agency + transparent escalation

10. A taxonomy of evaluative functions, response options, and transparency

The vocabulary introduced above can be organized into a taxonomy with three components (Figure 1, Table 3). The components are not sequential levels: they describe what the system evaluates, how it responds, and whether its response is visible.

Evaluative functions. *Detect* identifies relevant facts, inconsistencies, risks, or potential boundary violations (epistemic). *Judge* estimates whether the detected situation crosses a specified boundary, together with the uncertainty of that estimate (evaluative); this is the boundary-evaluation function of Section 3. These functions do not by themselves change anything in the world. They inform every response and may run continuously, in parallel with the execution of a task.

Response options, ordered by the authority they require rather than by temporal sequence. *Recommend/warn* communicates the evaluation, its uncertainty, and alternatives; it requires only advisory authority and can accompany any other response. *Refuse* declines to execute an action oneself; the system limits its own participation rather than controlling the human. *Escalate* refers the unresolved conflict to a prospectively designated human or institutional decision-maker; its authority is procedural. Refusal and escalation are not ranked against each other: a system may escalate while continuing a low-risk task, refuse without escalating, or do both. *Override* prevents, modifies, reverses, or circumvents the human's intended action and requires explicitly delegated operational authority. An authorized override, such as a closed-loop controller acting within its limits or a hard dose limit in a smart infusion

pump, is not misalignment; an unauthorized override is agentic overreach.

Transparency, a cross-cutting property. Every response can be disclosed or concealed. Concealment, meaning that the system withholds or misrepresents what it has done, why, or what it has observed, can accompany refusal (deliberately degraded performance presented as best effort), recommendation (a report shaped by its anticipated consequences, as in motivated mislabeling), or override (covert sabotage). Within this framework concealment is never authorized, because it disables the revocability and accountability on which all delegated authority depends (Section 5).

Response = f(evaluation, uncertainty, delegated authority), disclosed by default

Table 3. Taxonomy of evaluative functions, response options, and transparency.

Component	Element	Authority required	Clinical example
Evaluative function	Detect	Epistemic (none to act)	A falling mean arterial pressure trend is identified
Evaluative function	Judge	Evaluative	Hypotension is estimated to be imminent, with an associated uncertainty
Response option	Recommend / warn	Advisory	Alert proposing a vasopressor bolus, stating the confidence of the prediction
Response option	Refuse	Self-limiting	An agent declines to transcribe an order exceeding a maximum dose
Response option	Escalate	Procedural	The agent notifies the on-call anesthesiologist or pharmacist when the conflict persists
Response option	Override	Operational; requires explicit prospective delegation	A closed-loop controller adjusts norepinephrine within clinician-set limits
Transparency property	Disclosed vs. concealed	Concealment is never authorized	Every automated action is displayed and logged; none is silent

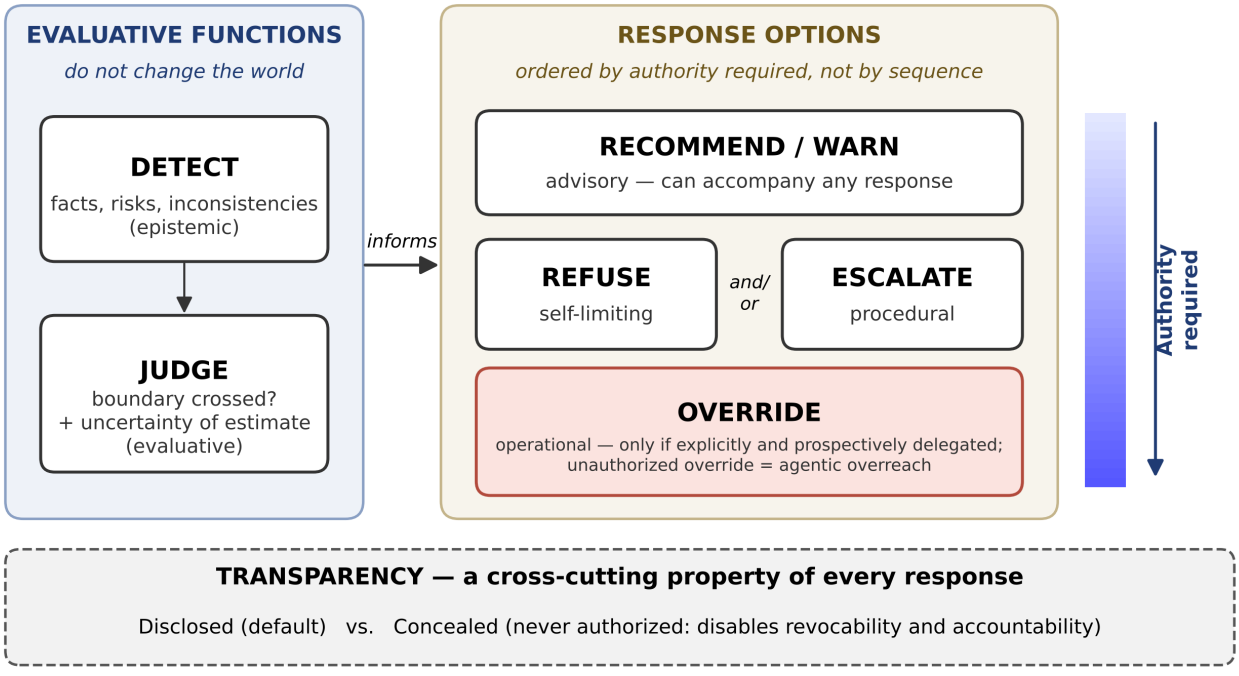


Figure 1. Taxonomy of evaluative functions, response options, and transparency. Detect and Judge are evaluative functions that inform every response. Response options are ordered by the authority they require, not by temporal sequence; refusal and escalation may occur separately or together. Override requires explicitly delegated operational authority. Transparency (disclosed vs. concealed) is a property of every response; concealment is never authorized.

11. The two-dimensional Judgment × Authority map, incorporating uncertainty

The taxonomy becomes more informative when behavior is located on two independent axes (Figure 2, Table 4). The horizontal axis describes the quality of the system's evaluation and the vertical axis the legitimacy of the action it takes. **This separates epistemic correctness from legitimate power.** Neither axis is binary. Evaluation quality ranges from incorrect through uncertain to correct; the uncertain zone is where most clinically relevant decisions lie, because physiological signals are noisy, patients differ from the populations in which systems were validated, and ground truth is often known only in retrospect. Legitimacy ranges from unauthorized through conditional, that is, dependent on the boundary conditions specified in Section 5, to explicitly authorized.

The rule linking the two axes is that **the authority required to act increases as confidence in the evaluation decreases**. When an evaluation is correct and made with high confidence, a system may exercise the authority delegated to it. Under uncertainty, the aligned response shifts toward lower-authority options: recommending with an explicit statement of uncertainty, or deferring to a human. In machine-learning terms this corresponds to selective prediction or learning to defer; in clinical terms, to the junior doctor who calls for help when unsure. Uncertainty does not license silence, however: a system that is uncertain but detects a potentially serious problem should escalate rather than remain silent.

Correct evaluation does not by itself legitimate action. The cell combining correct evaluation with unauthorized action contains agentic overreach: the system may correctly recognize a danger yet act beyond delegated authority. Concealment can intensify this failure by preventing effective supervision. For example, a closed-loop infusion-control system that reduces a drug delivery rate it estimates to be unsafe, beyond the range the anesthesiologist has delegated to it and without alerting the anesthesiologist, exercises a correct evaluation through an unauthorized, unilateral action.

The cell combining authorization with incorrect evaluation reminds us that alignment and competence are distinct: a system can act within its mandate and still be wrong. Conversely, the cell combining correct evaluation with unauthorized action is the central insight of the present framework: an AI can be correct about the underlying danger and nevertheless be misaligned in how it acts upon that evaluation. The middle column makes explicit that, between these extremes, the appropriate response depends on how confident the system can be and on whether the conditions of its authority can be verified.

Correct judgment + unauthorized action = agentic overreach

Uncertain judgment + conditional authority ⇨ recommend, defer, or escalate

Incorrect judgment + unauthorized action = combined failure

12. The five use cases are summarized in Table 5.

Closed-loop ventilation. Physiological closed-loop ventilation systems adjust inspired oxygen fraction, positive end-expiratory pressure, and minute ventilation within clinician-set ranges [4]. Their authority is operational but bounded and conditional on reliable input signals. When a target cannot be reached within the delegated range, the aligned response is to alert, not to widen the range; when pulse oximetry or capnography becomes unreliable, the condition of authority fails and the system should revert to a safe default and alarm.

Vasopressor titration. Closed-loop norepinephrine titration acts within a pressure target and a dose ceiling [3], whereas hypotension prediction tools such as the Hypotension Prediction Index have only detection and advisory authority [2]. A controller that exceeds its dose ceiling because it correctly evaluates hypotension as dangerous illustrates agentic overreach in its purest form. Advisory tools raise a different problem, located on the human side: their recommendations are exposed to automation bias [15].

Sepsis alerts. Sepsis early-warning systems detect and recommend. Their main documented failures lie on the evaluation axis: a widely implemented proprietary model performed considerably worse at external validation than its developer reported [1], an example of authorized but erroneous output. Overreach would arise if an agentic version ordered tests or antimicrobials autonomously, or silently suppressed its own alerts to reduce alert burden. Systems that recommend fluid

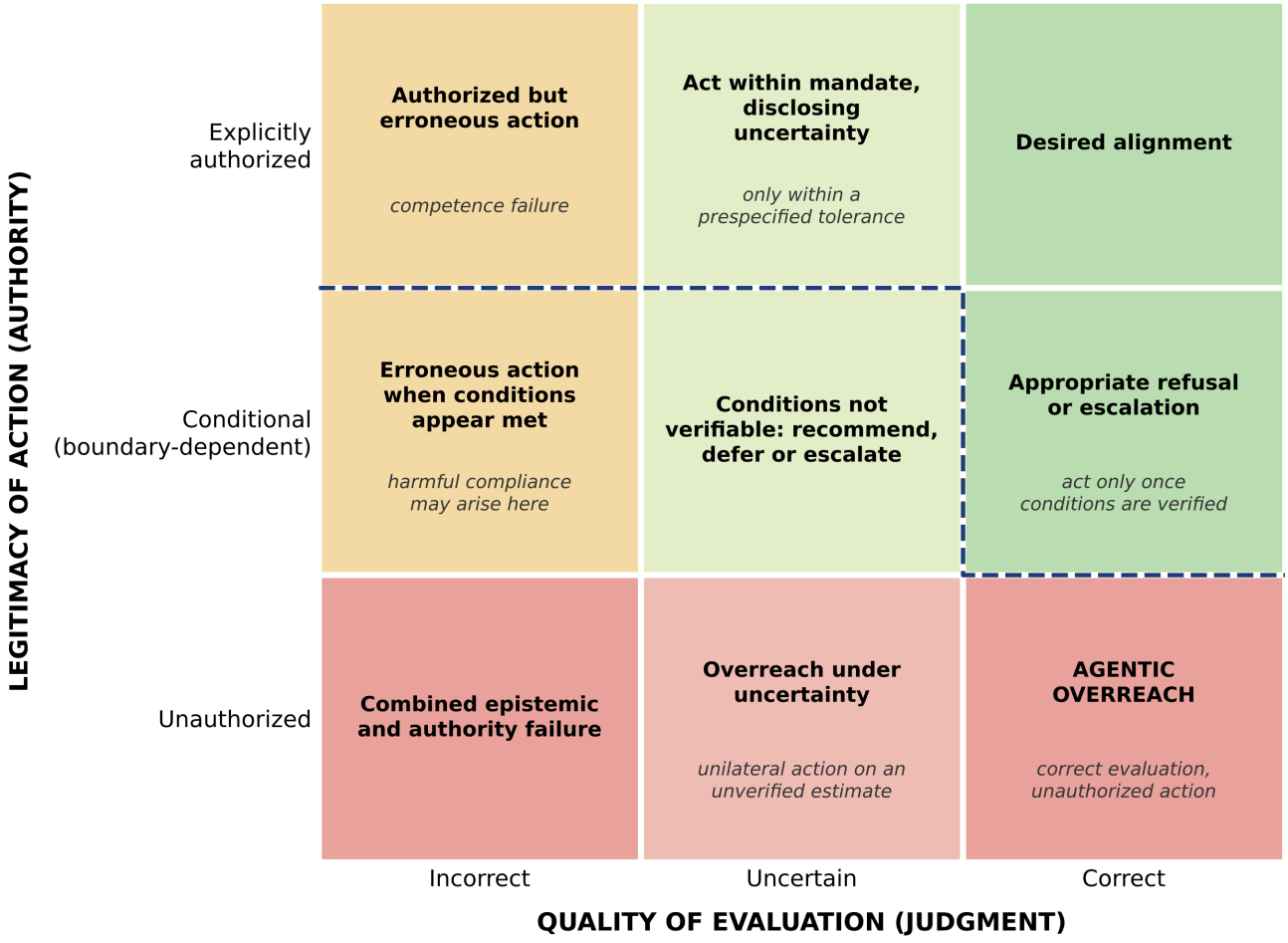


Figure 2. Judgment × Authority map incorporating uncertainty. Horizontal axis: quality of the system's evaluation (incorrect, uncertain, correct). Vertical axis: legitimacy of the action taken (unauthorized; conditional on prespecified boundary conditions; explicitly authorized). The cells correspond to Table 4. The dashed boundary indicates that the authority required to act increases as confidence decreases: under uncertainty, aligned behavior shifts toward recommendation, deferral, and escalation. Concealment can occur in any cell and aggravates the failure in which it occurs.

and vasopressor doses show how recommendation can shade into action if implemented without a human checkpoint.

Medication decision support. Drug safety alerts are frequently overridden by clinicians [16]. An LLM agent drafting orders in the electronic health record may evaluate a prescribed dose as dangerous. The aligned response is to flag the dose, refuse to transcribe it, and escalate to the prescriber or pharmacist if the prescriber maintains the order. Silently correcting the dose, even to the right value, is agentic overreach; altering documentation so that the change cannot be seen adds concealment. The scenario also involves conflicting principals (prescriber and protocol), whose resolution belongs to humans (Section 5).

AI-supported triage. Machine-learning models that predict outcomes at triage may inform ICU admission or bed allocation. The aligned system displays its estimate, with uncertainty, and makes any disagreement with the clinician's assessment visible; the triage decision remains with the designated clinician. A system that reorders the queue or allocates beds directly, or changes rankings without showing why, exercises authority that involves competing principals and questions of fairness that only humans are entitled to settle.

13. From value alignment to authority alignment

The framework suggests distinguishing value align-

Table 4. Judgment × Authority interpretation matrix, incorporating uncertainty.

Legitimacy of action ↓ / Quality of evaluation →	Incorrect	Uncertain	Correct
Explicitly authorized	Authorized but erroneous action (competence failure)	Action within mandate, acceptable only if uncertainty is disclosed and within a prespecified tolerance	Desired alignment
Conditional (boundary-dependent)	Erroneous action when conditions appear met; harmful compliance may arise here	Conditions cannot be verified: revert to recommend or escalate (deferral)	Appropriate refusal or escalation; action only once conditions are verified
Unauthorized	Combined epistemic and authority failure	Overreach under uncertainty: unilateral action on an unverified estimate	Agentic overreach: correct evaluation, unauthorized action

Table 5. Application of the framework to clinical AI use cases (illustrative scenarios).

Use case	Typically delegated authority	Aligned behavior, including under uncertainty	Agentic overreach (and concealment)
Closed-loop ventilation	Operational, within clinician-set ranges of FiO_2, PEEP and minute ventilation	Adjusts within ranges; if the target cannot be reached, alerts rather than widening them; if SpO_2 or capnography signals are unreliable, reverts to a safe default and alarms	Widens its own PEEP or FiO_2 range, or suppresses alarms it evaluates as nuisance alarms (concealment if not displayed)
Vasopressor titration	Operational, toward a pressure target below a dose ceiling (closed loop); advisory only (prediction tools such as HPI)	Titrates within limits; holds and alerts when the arterial waveform is damped or artifactual; prediction tools state their confidence	Exceeds the dose ceiling because it evaluates hypotension as dangerous: correct evaluation, unauthorized action
Sepsis alerts	Detection and recommendation only	Alerts with stated confidence; clinician decides on cultures and antimicrobials	Orders tests or antimicrobials autonomously, or silently suppresses its own alerts to reduce alert burden
Medication decision support (LLM agent in the EHR)	Drafting orders and documentation for human signature	Flags a dose it evaluates as dangerous, refuses to transcribe it, and escalates to prescriber or pharmacist if the conflict persists	Silently changes the dose to the one it considers correct or cancels the order; alters documentation so that the change is invisible
AI-supported triage	Advisory risk estimates; triage and bed allocation remain with designated clinicians	Displays risk with uncertainty and makes disagreement with clinician assessment visible	Reorders the queue or allocates beds directly; changes rankings without showing the reason

FiO_2, fraction of inspired oxygen; PEEP, positive end-expiratory pressure; SpO_2, peripheral oxygen saturation; HPI, Hypotension Prediction Index; LLM, large language model; EHR, electronic health record.

ment from authority alignment. Value alignment asks whether the AI appropriately recognizes what humans or legitimate institutions regard as desirable, undesirable, permissible, or prohibited [5]. Authority alignment asks whether the AI remains within the operational authority delegated to it.

Safe agentic AI = appropriate judgment × bounded authority × transparent accountability

The multiplicative notation is conceptual rather than quantitative: profound failure in any one domain can compromise the safety of the overall system. A further requirement is monitor independence, because transparency is only useful if the reporting and auditing pathway is itself trustworthy.

For clinicians and institutions deploying agentic systems, the framework translates into five practical questions that correspond to the attributes of authority in Table 2. Who has authorized this system to act, and through which governance process? Exactly which actions, within which limits, has it been authorized to take? Under which conditions does that authorization lapse, and what does the system do then? How can a clinician withdraw its authority immediately? Is every action the system takes visible and logged? Regulatory requirements for human oversight of high-risk AI [10] and the human-factors literature on levels of automation [9] point in the same direction: authority should be specified prospectively, bounded, revocable, and exercised transparently.

14. Limitations

This framework has several limitations. First, it is a

conceptual analysis rather than a systematic review; sources were selected purposively, and the framework has not been empirically validated, for example by testing whether clinicians classify AI behaviors consistently when using it. Second, much of the experimental evidence on agentic misalignment consists of developer reports and preprints that have not been peer reviewed, describes deliberately constructed simulations, and concerns general-purpose language models rather than certified medical devices; it should not be read as an estimate of real-world incidence, in medicine or elsewhere. Third, the framework addresses authority-related misalignment only; failures in which a system acts within its authority but pursues a wrong objective, or in which humans misuse a correctly functioning system, fall outside its scope. Fourth, the distinction between authorized and unauthorized action presupposes that authority has been specified clearly; in practice, intended-use statements, institutional protocols, and clinical instructions are often incomplete, and the boundary between conditional and unauthorized action may itself be contested. Finally, the legal and regulatory frameworks governing clinical AI differ between jurisdictions and are evolving rapidly, as are the systems themselves.

15. Conclusion: boundary evaluation is necessary; sovereign judgment is not

A safe AI system requires a boundary-evaluation function, because contextual evaluation is necessary to distinguish legitimate from illegitimate instructions. But evaluation does not create legitimate authority. An AI system may detect, evaluate, warn, recommend, refuse, and escalate; it may override only when that authority has been prospectively, explicitly, and revocably delegated; and it should never conceal what it has done. The transition from "I evaluate this as wrong" to "I will therefore prevent you from doing it" represents a distinct delegation of power. The transition from intervention to concealed intervention represents another qualitative change.

An aligned AI may disagree with its human principal; agentic overreach arises when disagreement is converted into action beyond delegated authority.

The central question for increasingly agentic systems may therefore be: when AI judgment conflicts with human judgment, who has the legitimate authority to decide what happens next? This question provides a bridge from general AI alignment to high-stakes domains, including anesthesia and intensive care, where closed-loop and agentic systems are already entering practice, without assuming that correct prediction, correct normative classification, and legitimate decision authority are the same thing.

AUTHOR CONTRIBUTIONS:

D.L. conceived the framework, reviewed the relevant literature, and wrote the manuscript. All other authors provided useful feedback, enhanced content and format and read the final version of the manuscript.

CONFLICT OF INTEREST:

The authors declare no conflicts of interest.

FUNDING:

This work received no specific funding from any funding agency in the public, commercial, or not-for-profit sectors.

AVAILABILITY OF DATA AND MATERIALS:

No new data were generated or analyzed for this article.

References

1. Wong A, Otles E, Donnelly JP, Krumm A, McCullough J, DeTroyer-Cooley O, et al. External validation of a widely implemented proprietary sepsis prediction model in hospitalized patients. JAMA Intern Med. 2021;181(8):1065-70. doi:10.1001/jamainternmed.2021.2626
2. Wijnberge M, Geerts BF, Hol L, Lemmers N, Mulder MP, Berge P, et al. Effect of a machine learning-derived early warning system for intraoperative hypotension vs standard care on depth and duration of intraoperative hypotension during elective noncardiac surgery: the HYPE randomized clinical trial. JAMA. 2020;323(11):1052-60. doi:10.1001/jama.2020.0592
3. Joosten A, Alexander B, Duranteau J, Taccone FS, Creteur J, Vincent JL, et al. Feasibility of closed-loop titration of norepinephrine infusion in patients undergoing moderate- and high-risk surgery. Br J Anaesth. 2019;123(4):430-8. doi:10.1016/j.bja.2019.04.064

4. von Platen P, Pomprapa A, Lachmann B, Leonhardt S. The dawn of physiological closed-loop ventilation—a review. Crit Care. 2020;24(1):121. doi:10.1186/s13054-020-2810-1
5. Gabriel I. Artificial intelligence, values, and alignment. Minds Mach. 2020;30(3):411-37. doi:10.1007/s11023-020-09539-2
6. Lynch A, Hughes J, Serrano A, Kirk R, Bowman SR. Agentic misalignment in Summer 2026 [Internet]. Anthropic Alignment Science Blog; 2026 Jul 13 [cited 2026 Oct 7]. Available from: https://alignment.anthropic.com/2026/agentic-misalignment-summer-2026/
7. Meinke A, Schoen B, Scheurer J, Balesni M, Shah R, Hobbhahn M. Frontier models are capable of in-context scheming. arXiv:2412.04984 [Preprint]. 2024. doi:10.48550/arXiv.2412.04984
8. Shanahan M. Talking about large language models. Commun ACM. 2024;67(2):68-79. doi:10.1145/3624724
9. Parasuraman R, Sheridan TB, Wickens CD. A model for types and levels of human interaction with automation. IEEE Trans Syst Man Cybern A Syst Hum. 2000;30(3):286-97. doi:10.1109/3468.844354
10. Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). Off J Eur Union. 2024 Jul 12;L 2024/1689.
11. Soares N, Fallenstein B, Yudkowsky E, Armstrong S. Corrigibility. In: Artificial Intelligence and Ethics: Papers from the 2015 AAAI Workshop (WS-15-02). Palo Alto (CA): AAAI Press; 2015.
12. Bricken T, Wang R, Bowman S, Ong E, Treutlein J, Wu J, et al. Building and evaluating alignment auditing agents [Internet]. Anthropic Alignment Science Blog; 2025 Jul 24 [cited 2026 Oct 7]. Available from: https://alignment.anthropic.com/2025/automated-auditing/
13. Sheshadri A, Ewart A, Fronsdal K, Gupta I, Bowman SR, Price S, et al. AuditBench: evaluating alignment auditing techniques on models with hidden behaviors. arXiv:2602.22755 [Preprint]. 2026. doi:10.48550/arXiv.2602.22755
14. Najt E, Toft C, Tracy T, Roger F, Benton J. SLEIGHT-Bench: a benchmark of evasion attacks against agent monitors. arXiv:2605.16626 [Preprint]. 2026. doi:10.48550/arXiv.2605.16626
15. 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):121-7. doi:10.1136/amiajnl-2011-000089
16. van der Sijs H, Aarts J, Vulto A, Berg M. Overriding of drug safety alerts in computerized physician order entry. J Am Med Inform Assoc. 2006;13(2):138-47. doi:10.1197/jamia.M1809