Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

How Much Should Artificial Intelligence Do Without a Physician? An Autonomy Scale for AI Agents in Solid Organ Transplantation.

Malik Muhammad Makki

dr.malikmakki@gmail.com

Augustana University https://orcid.org/0009-0003-3291-9150

Samia Yunas

Johns Hopkins Hospital https://orcid.org/0009-0004-1682-444X

Research Article

Keywords: agentic artificial intelligence, transplant governance, autonomy levels, clinical decision support, solid organ transplantation, regulatory science

Posted Date: November 11th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11691366/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

How Much Should Artificial Intelligence Do Without a Physician?

An Autonomy Scale for AI Agents in Solid Organ Transplantation.

Running title: Tiered oversight for transplant AI

Authors: Malik Muhammad Makki; Samia Yunas

[1]Medical College of Georgia, Augusta University, Augusta, GA, USA.

[2]Samia Yunas: Department of Transplant Nephrology, Johns Hopkins Hospital, Baltimore, MD, USA.

Corresponding author:

Malik Muhammad Makki MD, MBA
Medical College of Georgia, Augusta University, Augusta, GA, USA.
Email: dr.malikmakki@gmail.com; phone: 786-510-8761.

Structured abstract

Background: Artificial intelligence in surgery and transplantation is moving from tools that give a single output toward software agents that retrieve data, plan several steps, use other tools, and take actions. Existing autonomy classifications and device frameworks address physical autonomy and fixed-function outputs well, but they do not specify oversight for systems whose behavior depends on intermediate decisions and tool use.

Methods: We developed a governance framework by combining Yang and colleagues' six-level autonomy scale with an agentic overlay covering tool use, data access, action scope, and persistence. We mapped the framework against current U.S. guidance for AI-enabled devices and the American Medical Association's 2026 framework for agentic AI. We illustrated its application with four worked examples from solid organ transplantation: donor procurement, machine perfusion, and post-transplant monitoring with cell-free DNA and donor-specific antibodies, assigning each plausible agent function a level and governance tier and identifying the controls that apply.

Results: The framework assigns each agent function to one of four governance tiers based on autonomy level, reversibility, and consequence. Tier A (levels 0 to 1) requires accuracy testing and query logging. Tier B (level 2) requires bounded limits, full action logs, and drift monitoring. Tier C (level 3) requires clinician approval of each consequential action, prospective evaluation, and named accountability. Tier D (levels 4 to 5) is not recommended for high-consequence decisions in the current evidence period. In the transplant example, the organ-acceptance decision is assigned to Tier C with clinician decision authority, allocation pairing is excluded from agent action, and a protocol-defined perfusion-time limit is enforced deterministically outside the model. Across the four examples, clinicians approve

each consequential action, no viability marker drives acceptance alone, and automatic immunosuppression changes are excluded.

Conclusions: Oversight of agents in transplant care should scale with autonomy, reversibility, and consequence rather than with product category. Binding rules should be enforced in deterministic code rather than left to model reasoning. Prospective evaluation is required before any agent in a high-consequence transplant decision advances beyond Tier C.

Keywords: agentic artificial intelligence; transplant governance; autonomy levels; clinical decision support; solid organ transplantation; regulatory science

1. Introduction

Surgery is moving from tools that respond to a surgeon's hand toward systems that plan, retrieve information, and act across several steps. Robotic platforms already automate specific tasks under surgeon supervision, and the field has long used graded frameworks to describe how much a robot does on its own.[1] A newer class of artificial intelligence systems, often built on large language models, can chain functions together, call external tools, and execute actions with limited prompting.[4]

Figure 1 shows the six autonomy levels, the agentic overlay they are paired with, and the governance tier each level maps to. Table 1 gives the full definitions.

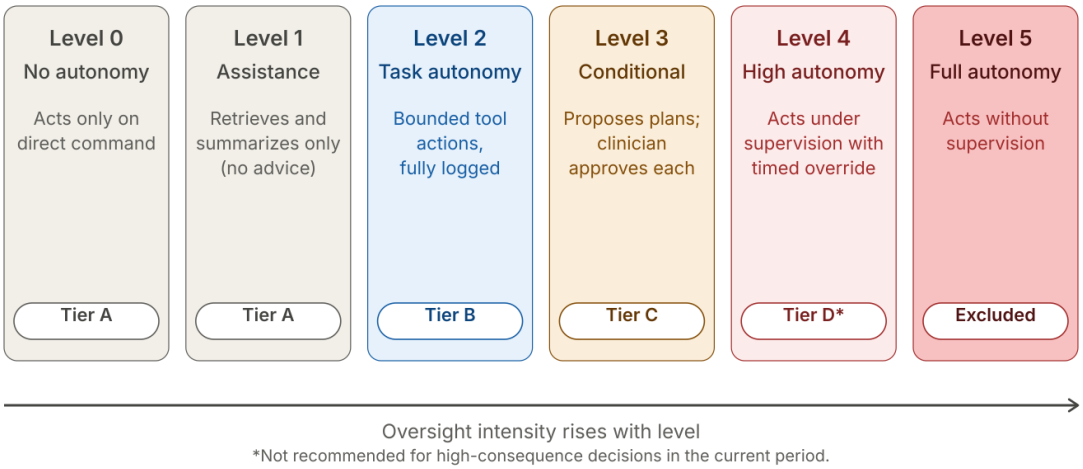


Figure 1. Six autonomy levels with governance tier. Level 4 is not recommended for high-consequence decisions in the current period, and level 5 is excluded.

Evidence on how much autonomy existing surgical devices are limited but informative. A 2024 systematic review of 49 FDA-cleared surgical robots, cleared between 2015 and 2023, classified each device on a five-level autonomy scale that follows the same logic as Yang et al.[10] The review reports that 86% of devices fell at Level 1, robot assistance, and 6% reached Level 3, conditional autonomy.[10]

Governance frameworks developed for surgical robotics and for AI-enabled devices were built around fixed functions and bounded outputs. The U.S. Food and Drug Administration's 2024 final guidance on predetermined change control plans allows manufacturers to pre-specify modifications to AI-enabled device software functions, which addresses post-market change in learning systems.[2] It does not address systems that choose their own sequence of retrieval and action at the point of care. The AMA's 2026 framework starts from the same concern: clinical authority should be earned through evidence and maintained through demonstrated performance, and what an AI system may do should be set explicitly.[3]

Surgical settings differ in how quickly decisions must be made, how reversible they are, and how far their consequences extend. Some agent functions, such as summarizing a patient's record, carry little risk. Others, such as recommending whether to proceed with an operation or how to allocate a scarce resource, carry substantial risk. A governance framework must distinguish between these cases rather than treating all clinical AI alike.

Solid organ transplantation illustrates the problem. Donor evaluation, allocation, and time-critical decisions operate under tight timelines and under rules that link individual decisions to population outcomes. Transplant also depends on registry data and policies that change over time. An agent that draws on a previous data version, or that encodes a superseded rule, may drift from current practice without anyone noticing. We therefore apply the framework to solid organ transplantation, where these demands are concentrated.

Early transplant-specific work is emerging. A recent study describes a transplant-specific large language model agent for post-transplant self-management.[6] A 2026 symposium report identified bias, explainability, and acceptability as key barriers to meaningful use in transplant.[5] To our knowledge, none of these sets out governance requirements that vary by autonomy level for agentic systems.

Figure 2 summarizes how these developments have unfolded, from the autonomy scale for robotics through recent agent-specific guidance and transplant policy.

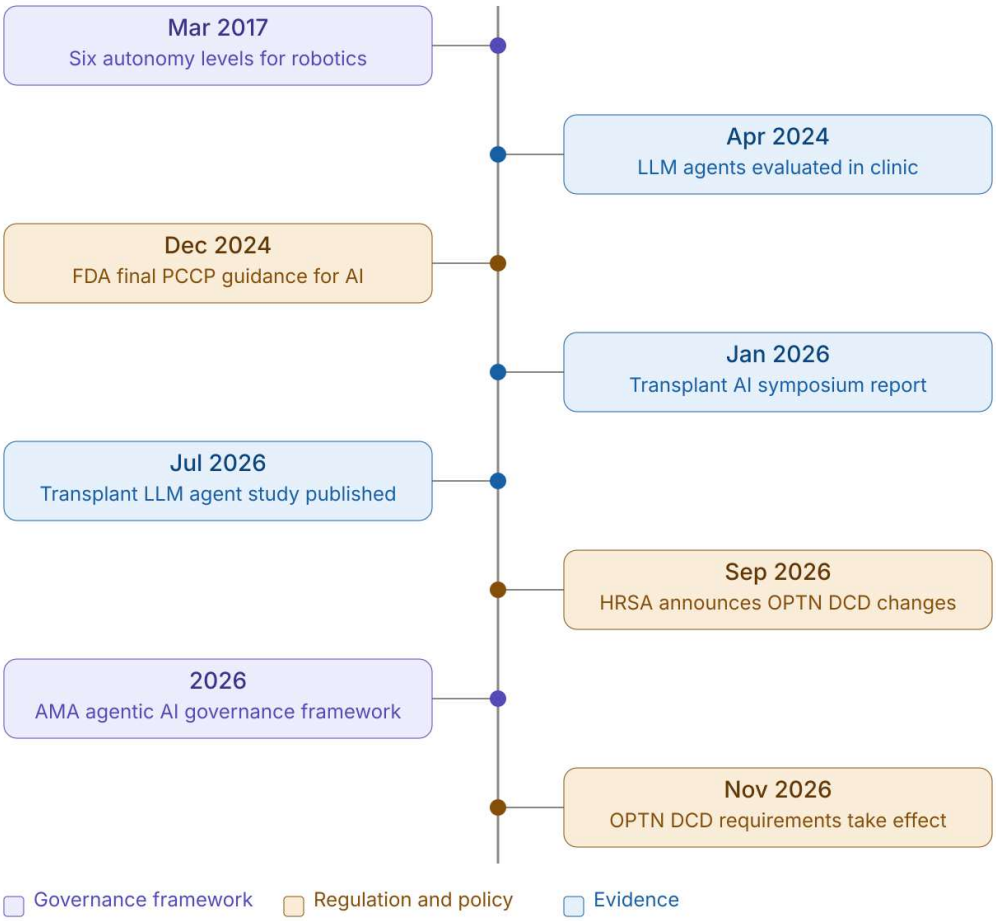


Figure 2. Timeline of governance and evidence milestones, March 2017 to November 2026, colored by type: governance framework, regulation and policy, and evidence.

We make three contributions.

First, we extend an established six-level autonomy scale with an agentic overlay that describes tool use and multi-step action.

Second, we propose a four-tier governance structure that assigns oversight requirements by autonomy level, reversibility, and consequence.

Third, we apply the framework to four solid organ transplant examples: donor procurement, machine perfusion, and post-transplant monitoring with cell-free DNA and with donor-specific antibodies. These show how the tiers apply in practice and where binding rules must be enforced outside the model.

2. Methods

2.1 Framework development

We began with Yang and colleagues' six-level autonomy scale for medical robotics, which defines levels from no autonomy (level 0) to full autonomy (level 5).[1] We retained these levels because they are widely cited and map onto existing regulatory vocabulary. We then added an agentic overlay with four operational dimensions: tool use, data access, action scope, and persistence across sessions.

We drew on the FDA final guidance on predetermined change control plans for AI-enabled device software functions (December 2024);[2] the AMA governance framework for agentic AI assuming clinical functions (2026);[3] and published frameworks for evaluating and governing large language model agents in clinical settings.[4,8,9] These sources were selected for direct relevance to governance of AI-enabled devices and clinical agents. This was a targeted narrative review rather than a systematic search. The governance tiers are a proposal derived from Yang and colleagues' levels and from reversibility and consequence criteria. They are not an adopted standard and have not been validated by an expert panel.

2.2 Worked examples

We used four worked examples from solid organ transplantation, each from a different stage of care: donor procurement, machine perfusion, and post-transplant monitoring with cell-free DNA and with donor-specific antibodies. Each example describes plausible agent functions rather than a deployed system. For each function we assigned an autonomy level and tier and identified the controls that apply. The assignments are proposals for review.

Procurement stages were identified from standard sequences. Machine perfusion covers normothermic machine perfusion (NMP) and hypothermic oxygenated perfusion (HOPE). The two monitoring examples cover cell-free DNA (cfDNA) and donor-specific antibodies (DSA) separately. No binding national OPTN requirement for machine perfusion time was identified, so the hard constraint in the machine perfusion example is a protocol rule set by the institution rather than a policy requirement.

2.3 Ethics

This study does not involve human participants and used no identifiable patient data.

3. Results

3.1 Autonomy levels

Table 1. Autonomy levels with agentic overlay and governance tier

Level	Definition (adapted from Yang et al.[1])	Agentic overlay	Governance tier	Example
0. No autonomy	Device or system acts only on direct user command	No AI-initiated action or retrieval	A	Manual record review without AI
1. Assistance	Provides information or mechanical support under direct control	Retrieves and summarizes; no recommendation	A	Summarizing a patient's recorded history
2. Task autonomy	Performs specified tasks autonomously when initiated by a human	Executes bounded, reversible tool actions with full logging	B	Drafting monitoring alerts for clinician review
3. Conditional autonomy	Generates task strategies; a human selects or approves them	Proposes multi-step plans and recommendations requiring approval	C	Recommending an operative plan for surgeon approval
4. High autonomy	Makes medical decisions under clinician supervision	Acts on decisions with supervision and a timed override	D (not recommended for high-consequence decisions now)	Intraoperative decision with timed override
5. Full autonomy	Performs an entire task or procedure without supervision	Not applicable to current clinical practice	Excluded	Not applicable

Tier A: minimum controls. **Tier B:** bounded limits and monitoring. **Tier C:** clinician approval of each consequential action, prospective evaluation, and named accountability. **Tier D**: prospective trials only.

Three design principles govern use of the table. Level is set by function, not by product, so one system may operate at different levels for different functions. Level is set by reversibility and consequence, so a function that cannot be reversed, or whose override window is shorter than clinicians can reasonably respond to, should not operate above level 2 regardless of accuracy. Changes in level, data sources, or governing policy trigger revalidation.

Figure 3 gives the decision procedure for assigning a function to a tier.

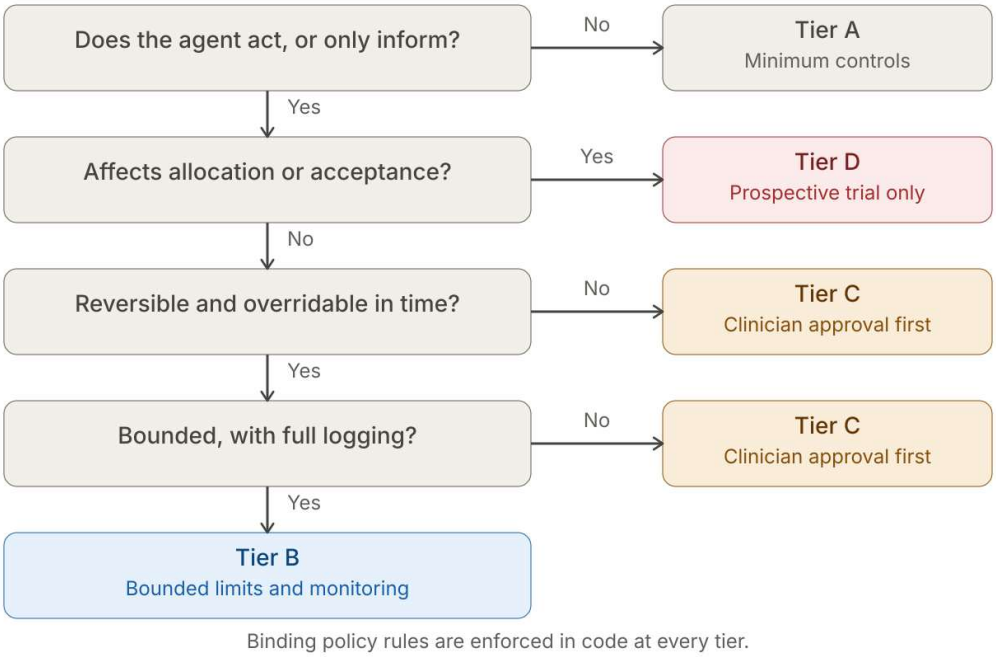


Figure 3. Decision tree for assigning a governance tier. Each function passes through four questions in sequence, and each answer routes it to a tier. Binding policy rules are enforced in code at every tier.

3.2 Risk domains

Table 2. Risk domains for agentic clinical AI

Domain	Mechanism	Minimum control
Compounding error	Small errors at one step propagate through later tool calls	Step-level logging and checkpoint validation
Data provenance and drift	Clinical data and policies are revised; retrieved sources change	Versioned data and policy sources; revalidation on each change
Opacity	Multi-step reasoning is difficult to inspect after the fact	Retained reasoning traces and action logs
Automation bias	Clinicians accept agent outputs under time pressure	Independent review at tiers A and C; measured override rates
Accountability diffusion	An action spans vendor, institution, and clinician	Named responsibility for each function and tier
Policy violation	An agent acts against a binding rule it has not been given or has misapplied	Deterministic enforcement of hard constraints outside the model
Equity	Errors in high-consequence decisions fall unevenly across populations	Subgroup performance reporting before and after deployment

Continuous governance is distinct from one-time evaluation. A recent governance framework for an electronic health record-embedded clinical agent treats logging, live

clinician feedback, and controlled experimentation before deployment as ongoing requirements rather than pre-deployment steps.[8]

Figure 4 shows the operating loop that the framework governs.

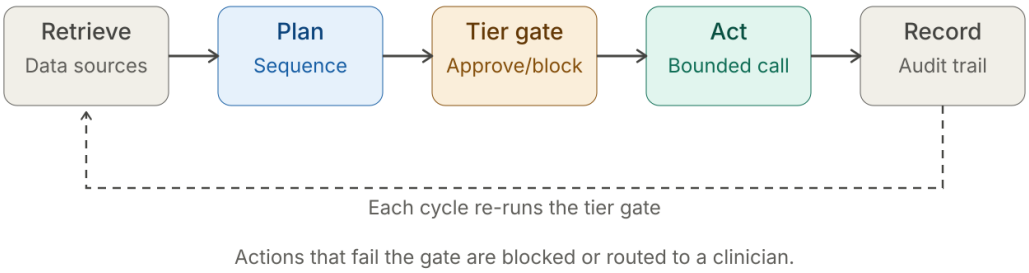


Figure 4. Agent operating loop with a tier gate before action. The loop retrieves data, plans, passes the tier gate, acts through a bounded tool call, and records the action in the audit trail. Each cycle re-runs the gate.

Figure 5 shows the relative priority of controls across risk domains and tiers. The ratings are illustrative, derived from Tables 2 and 3, and are not empirical estimates.

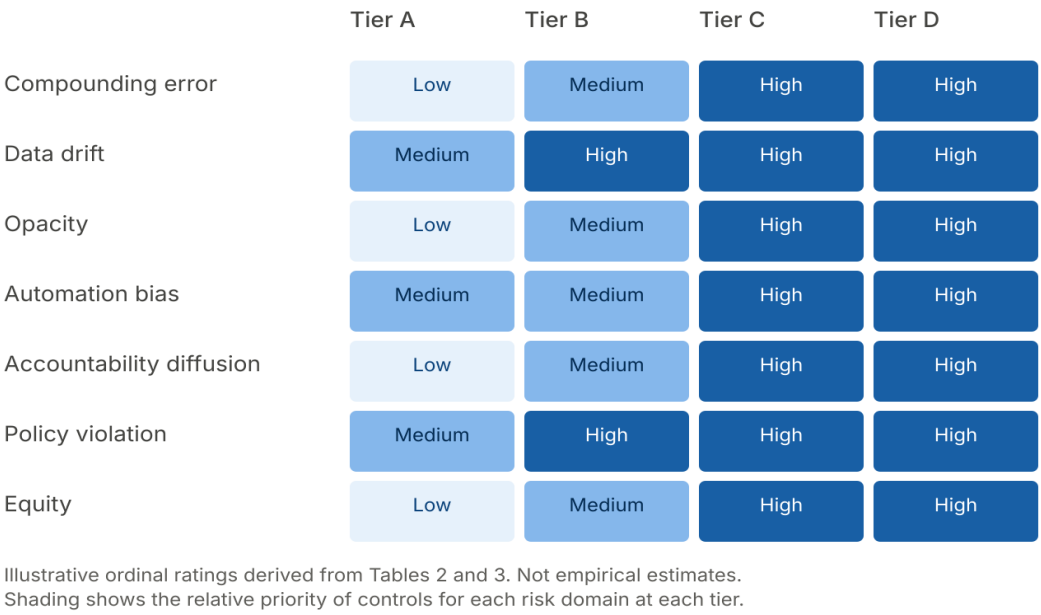


Figure 5. Relative priority of risk controls by governance tier. Shading shows the priority of controls for each risk domain at each tier.

3.3 Required controls by tier

Table 3. Required controls by governance tier

Control	Tier A (levels 0 to 1)	Tier B (level 2)	Tier C (level 3)	Tier D (levels 4 to 5)
Pre-deployment evaluation	Accuracy against clinician decisions	Scenario and adversarial testing	Prospective evaluation	Prospective trial only
Action audit	Query log	Full action log	Full action log with approval record	Full log with independent review
Human role	Reads and decides	Sets limits; reviews logs	Approves each consequential action	Supervises with timed override
Monitoring	Annual review	Monthly drift check	Continuous monitoring	Continuous monitoring with external audit
Revalidation trigger	Data source change	Data, model, or policy change	Any change in function, limits, or policy	Not applicable
Accountable owner	Institution	Institution and named clinician	Institution, vendor, and named clinician	Not applicable

Hard constraints. Binding rules, such as a protocol-set perfusion-time limit or a required disclosure step, should be enforced by deterministic code outside the model. The agent should be unable to override them through its reasoning. Each such rule carries its policy version, and any policy change triggers revalidation.

3.4 Worked example 1: donor procurement

Table 4 assigns autonomy levels across the procurement stages. Figure 6 shows the same stages with their tier assignments.

Table 4. Autonomy levels across donor procurement (worked example 1)

Pathway stage	Plausible agent function	Level	Tier	Controls and constraints
Donor record review	Summarize medical history and prior testing	1	A	Version-controlled sources; query log
Risk estimation	Estimate outcome risk from donor predictors	1	A	Recalibration on each new data release; subgroup reporting
Organ-acceptance decision	Summarize donor and perfusion data and flag deviations	3 at most	C	Clinician decides; no autonomous action on the decision
Allocation pairing	Propose recipient matching	Not assigned	D	Excluded for current period; equity review required

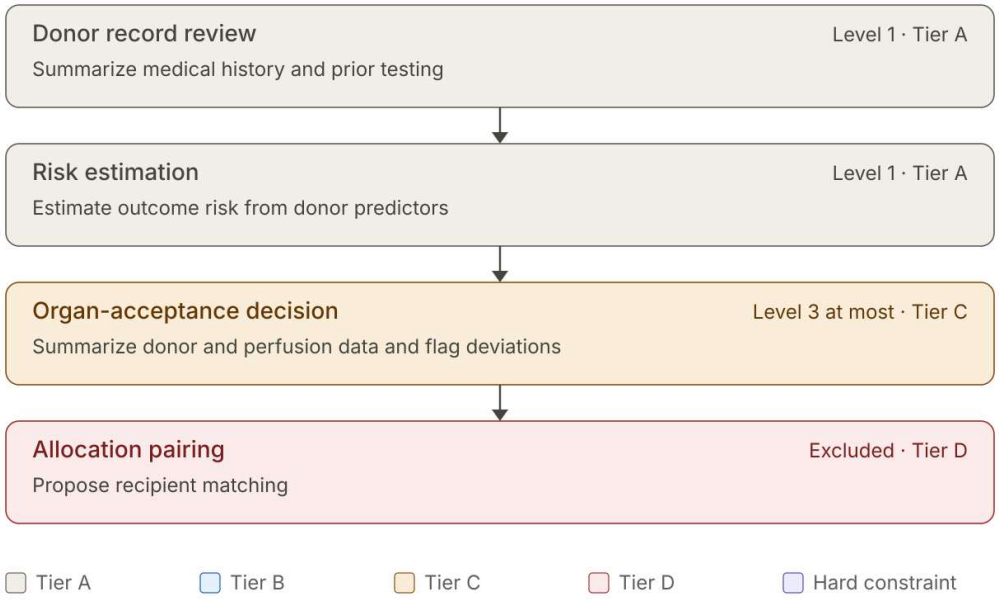


Figure 6. Donor procurement stages with autonomy level and governance tier (worked example 1).

The organ-acceptance decision is the most consequential stage in this example. An agent can support it by summarizing donor and perfusion data and flagging deviations, but the decision remains with the clinician.

3.5 Worked example 2: machine perfusion

Table 5 assigns autonomy levels across normothermic and hypothermic oxygenated machine perfusion.

Table 5. Autonomy levels across machine perfusion (worked example 2)

Function	Plausible agent function	Level	Tier	Controls and constraints
Perfusion-duration planning	Suggest a target perfusion duration for NMP or HOPE	3	C	Clinician approval; logged rationale; protocol limits enforced in code
Protocol monitoring	Adjust perfusion monitoring alerts	2	B	Bounded thresholds; full action log; override tracking
Perfusion summary	Summarize perfusion parameters over time	1	A	Version-controlled device data; query log
Maximum perfusion time	Track elapsed perfusion time against the protocol limit	Not applicable (rule, not agent)	Hard constraint	Deterministic timer outside the model; protocol version recorded

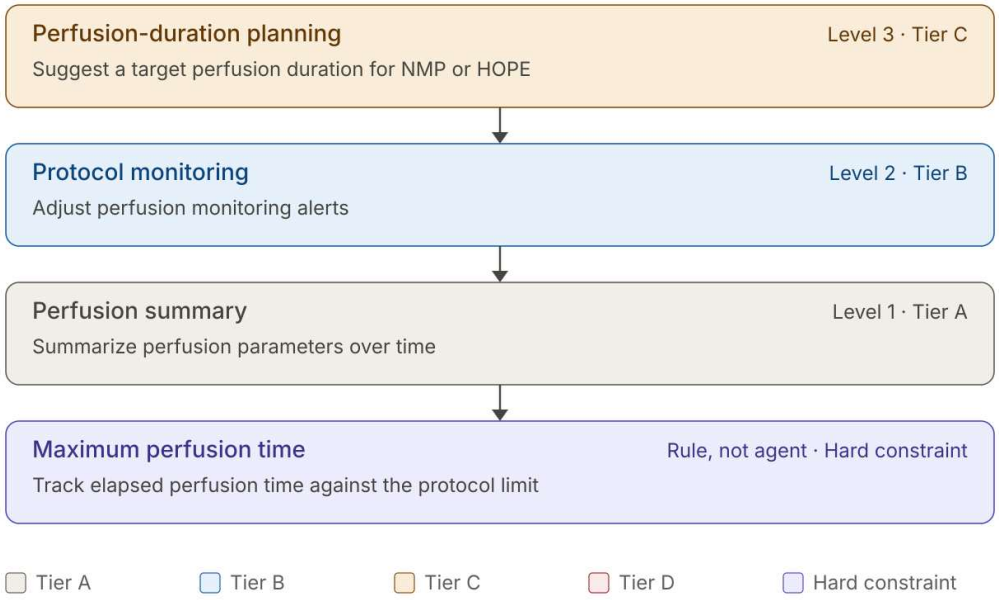


Figure 7. Machine perfusion stages with autonomy level and governance tier (worked example 2).

The maximum perfusion time is a binding rule for this example. It should be enforced in deterministic code, and the agent should not be able to bypass it or reason its way around it. Protocol limits are set separately for NMP and HOPE.

3.6 Worked example 3: cell-free DNA

Table 6 assigns autonomy levels across post-transplant monitoring with cfDNA.

Table 6. Autonomy levels across cell-free DNA monitoring (worked example 3)

Function	Plausible agent function	Level	Tier	Controls and constraints
cfDNA summary	Summarize cfDNA results over time	1	A	Versioned laboratory sources; query log
Injury estimate	Estimate graft injury risk from cfDNA trends	3	C	Clinician review; prospective evaluation
cfDNA alert	Flag a rise in cfDNA above a validated threshold	2	B	Threshold set by validation study; alert log
Automatic clinical action	Act on a cfDNA result without clinician approval	Not assigned	D	Excluded

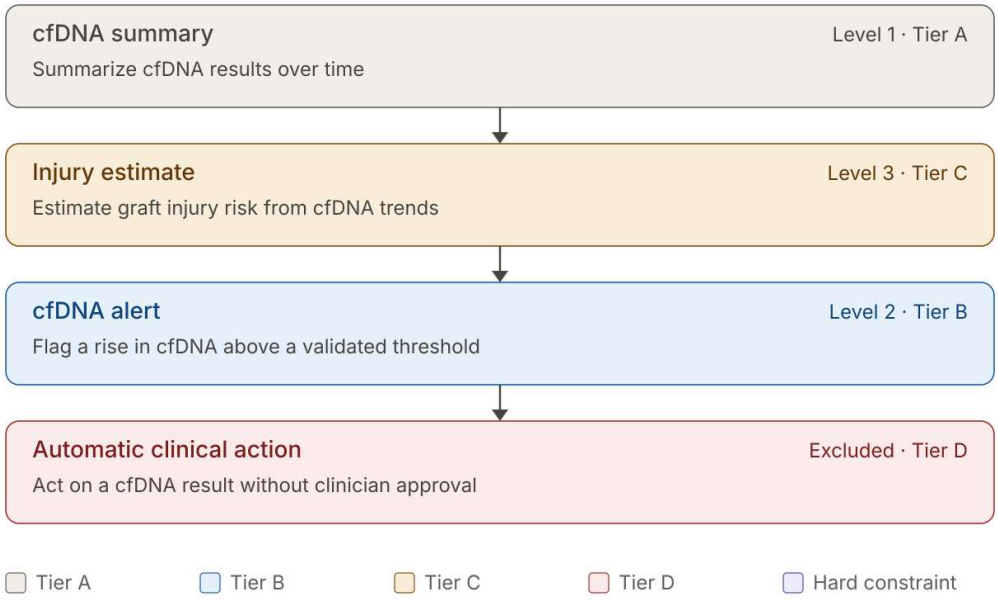


Figure 8. Cell-free DNA monitoring stages with autonomy level and governance tier (worked example 3).

A cfDNA result should inform clinical review. A single value should not drive a clinical decision until prospective evaluation supports it.

3.7 Worked example 4: donor-specific antibodies

Table 7 assigns autonomy levels across post-transplant monitoring with DSA.

Table 7. Autonomy levels across donor-specific antibody monitoring (worked example 4)

Function	Plausible agent function	Level	Tier	Controls and constraints
DSA summary	Summarize donor-specific antibody results over time	1	A	Versioned laboratory sources; query log
New DSA alert	Flag a newly detected DSA or rising titer	2	B	Threshold set by validation study; alert log
Immunosuppression change	Recommend a change in immunosuppression	3	C	Clinician approves each change; logged rationale
Automatic change	Adjust immunosuppression without approval	Not assigned	D	Excluded

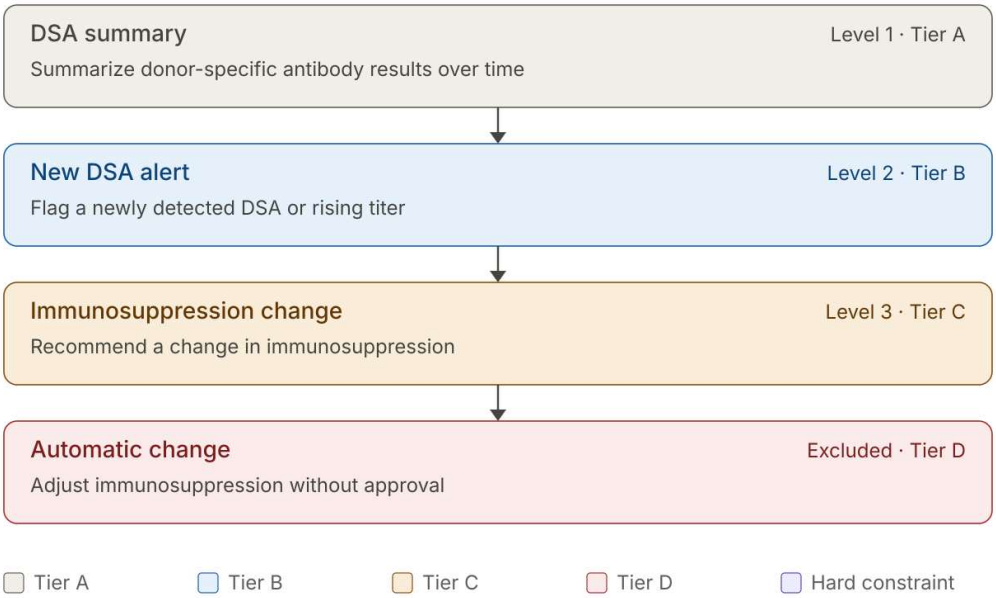


Figure 9. Donor-specific antibody monitoring stages with autonomy level and governance tier (worked example 4).

Immunosuppression decisions carry the highest consequence in this example. Recommendations remain at Tier C, and automatic changes are excluded.

4. Discussion

For clinicians, the framework answers one practical question: how much should an AI tool do without you? Level 0 and level 1 tools retrieve and summarize, and the clinician makes every decision. Level 2 tools may carry out bounded, reversible tasks, such as drafting an alert, but every action is logged and limited. Level 3 tools may propose a plan or recommendation, and nothing happens until the clinician approves it. Level 4 and level 5 tools, which would decide or act without supervision, are not recommended for high-consequence decisions in the current period. Separately, some limits, such as a maximum perfusion time, should be enforced by software the agent cannot override. In the transplant examples, the organ-acceptance decision stays with the clinician, and no single viability marker or laboratory value triggers an acceptance or decline on its own.

4.1 Relation to the AMA framework

The AMA framework and this paper share a central principle: clinical authority should be earned through evidence and maintained through demonstrated performance.[3] The AMA also states that greater consequence and greater autonomy call for stronger evidence, which is the basis for our tiers. The AMA organizes its work around four questions: what clinical function the AI is performing; what the AI should be permitted to do, and under what

conditions; what demonstrates that it can perform that function, and what happens if it fails; and who is responsible for the risks and outcomes.[3]

Our contribution is to make the second question, what the AI may do and under what conditions, operational. The autonomy levels and tier assignments specify which functions can be performed at what oversight, and the hard-constraint principle specifies what the system may never do, regardless of its evidence. We see the two frameworks as complementary. The AMA framework states the questions, and this paper proposes an implementation structure for answering them.

4.2 Relation to FDA regulation

The FDA's predetermined change control plan guidance addresses how manufacturers can pre-specify modifications to AI-enabled device software.[2] It is designed for devices whose function is defined at authorization. An agent that chooses its sequence of retrieval and action at run time presents a different question. Pre-specifying the set of permitted actions and their limits, as our tiers require, is one possible bridge between the two. Whether regulators would treat such limits as a predetermined change control plan is an open question.

A related concern is predicate creep. If most surgical devices reach the market through substantial-equivalence pathways, a manufacturer may add autonomous features while describing the device in terms of an existing human-controlled predicate. That could keep a device within a lower-scrutiny pathway even as its functions change.[10] Agentic systems raise the same question more sharply, because their functions can change through software updates in ways a fixed predicate does not capture. The requirements in Table 3 would apply regardless of the regulatory pathway.

4.3 Transferability beyond transplant

The taxonomy and controls are defined by function, reversibility, and consequence rather than by organ system, so they may transfer to other surgical settings, such as intraoperative decision support or perioperative monitoring. We have not tested this. Transfer would require new tier assignments for each setting, because the same function can carry different consequences in different procedures.

4.4 Why transplant is a demanding setting

Transplant combines three features that make governance harder than in many surgical settings. Allocation and donor-acceptance decisions carry population-level consequences. Time-critical decisions must be made under pressure, which raises the risk of automation bias. Binding policies change on fixed timelines, so an agent must be revalidated when rules change, not only when code changes. For example, OPTN policy on donation after circulatory death takes effect on November 2026 and new place kept updating.[7]

Clinical evidence in this area is still developing, and the evidence an agent relies on will change as new results appear. A governance framework should therefore require that this evidence be versioned and re-examined over time.

4.5 Equity

Agents that influence high-consequence decisions can affect populations unevenly even when overall accuracy is high. Subgroup performance reporting should be a precondition for any function above tier A. Prior work on bias in transplant artificial intelligence identifies this as a core barrier.[5]

4.6 Viability markers

Viability markers offer a different kind of agent function. Perfusate flavin mononucleotide (FMN) has been studied as a candidate biomarker of liver injury during normothermic machine perfusion,[11] but it has not been validated as a marker of viability. An agent could summarize FMN trends or estimate injury risk from them. Any such estimate remains a recommendation for clinician review at Tier C. A marker should not drive acceptance on its own until prospective evaluation shows that it adds to clinician judgment, so an acceptance recommendation based on a marker alone is excluded from agent action.

4.7 Relation to related preprints

Several recent preprints address governance of agentic systems from related angles. Ye and colleagues propose evaluating medical large language models by levels of autonomy, moving from benchmarks to applications.[9] Dhar and colleagues describe a pattern catalogue and orchestration framework for hospital information systems in which risk tiers govern runtime control.[12] Cuneo and colleagues present AI-GRACE, an operational framework that traces organizational obligations to deployment capabilities and architecture.[13] Naik and colleagues describe TRACCIA, a governance platform that records agent telemetry using OpenTelemetry.[14] These works concentrate on evaluation, system architecture, operational process, or telemetry. Our contribution differs in three respects. It links autonomy level, reversibility, and consequence in a single tier assignment; it specifies the controls required at each tier; and it applies these to surgical functions, with binding rules enforced in deterministic code. The approaches are complementary. Telemetry and runtime control can supply the evidence that the tier requirements in Table 3 call for.

5. Limitations

The framework was developed and illustrated in transplantation only. No agent in transplant or surgery has yet been evaluated prospectively at scale, so tier assignments reflect present evidence and accountability rather than demonstrated performance. The framework is a proposal. It has not been validated through a formal expert consensus process, in which a structured panel of surgeons, transplant clinicians, and ethicists would rate and revise the tier assignments. The worked example is illustrative and does not describe a deployed system. The autonomy levels were developed for robotics and extended here; their applicability to purely software agents requires testing. The regulatory and policy landscape is changing, and the framework reflects guidance and policy in effect at the time of writing.

6. Conclusions

Agentic AI will reach transplant care whether or not governance keeps pace. Oversight that scales with autonomy, reversibility, and consequence allows useful assistance while keeping clinicians responsible for decisions that matter most. Binding rules should be enforced in code rather than by model reasoning. In high-consequence settings such as solid organ transplantation, these requirements are needed now.

Declarations

Funding: No funding was received.

Conflicts of interest: The authors declare no conflicts of interest.

Author contributions: Both authors contributed to the conception, drafting, and revision of the manuscript and approved the final version.

Data availability: No new data were generated.

Use of AI tools: AI tools (a large language model) were used for general writing support during drafting. The authors reviewed and edited all content and take full responsibility for it.

References

1. Yang GZ, Cambias J, Cleary K, Daimler E, Drake J, Dupont PE, et al. Medical robotics: regulatory, ethical, and legal considerations for increasing levels of autonomy. Sci Robot. 2017;2:eaam8638. doi:10.1126/scirobotics.aam8638.

2. U.S. Food and Drug Administration. Marketing submission recommendations for a predetermined change control plan for artificial intelligence-enabled device software functions: final guidance. December 2024. Available from: https://www.fda.gov/media/187905/download

3. American Medical Association. A governance framework for agentic AI assuming clinical functions. Chicago: AMA Center for Digital Health and AI; 2026. Available from: https://www.ama-assn.org/practice-management/digital-health/governance-framework-agentic-ai-assuming-clinical-functions

4. Mehandru N, Miao BY, Rodriguez Almaraz E, Sushil M, Butte AJ, Alaa A. Evaluating large language models as agents in the clinic. NPJ Digit Med. 2024. doi:10.1038/s41746-024-01083-y.

5. Koivu A, Azarfar G, Maleki S, Sidhu A, Bhat M. 2024 transplant AI symposium: key AI models shaping the future of transplant care. Front Transplant. 2026;4:1723799. doi:10.3389/frtra.2025.1723799.

6. Zeng C, Zhou X, Xu HY, Zhu XP, Shi Y, Dai GL, et al. Real-world analysis of organ transplantation-specific agent based on large language model in post-transplant self-management during off-hours: a mixed-methods study. Curr Med Sci. 2026;46(4):1004-1013. doi:10.1007/s11596-026-00219-3.

7. U.S. Health Resources and Services Administration. Upcoming donation after circulatory death policy changes [OPTN bulletin]. September 18, 2026. Policy effective November 18, 2026. Available from: https://www.asts.org/connect/news/2026/09/24/optn-update–upcoming-donation-after-circulatory-death-policy-changes (reposting the OPTN bulletin).

8. Shah A, Hines A, Downs A, Bajet D, Mui P, Araujo F, et al. End-to-end evaluation and governance of an EHR-embedded AI agent for clinicians. arXiv:2604.27309. 2026. https://arxiv.org/abs/2604.27309

9. Ye X, Dineen J, Li Z, Xu Z, Chen W, Lu S, et al. Evaluating medical LLMs by levels of autonomy: a survey moving from benchmarks to applications. arXiv:2510.17764. 2025. https://arxiv.org/abs/2510.17764

10. Lee A, Baker TS, Bederson JB, Rapoport BI. Levels of autonomy in FDA-cleared surgical robots: a systematic review. NPJ Digit Med. 2024;7:103. doi:10.1038/s41746-024-01102-y.

11. Kesseli SJ, Gloria JN, Abraham N, Halpern SE, Cywinska GN, Zhang M, Moris D, Schmitz R, Shaw BI, Fitch ZW, Song M, Guy CD, Hartwig MG, Knechtle S, Barbas AS. Point-of-care assessment of DCD livers during normothermic machine perfusion in a nonhuman primate model. Hepatol Commun. 2021;5(9):1527-1542. doi:10.1002/hep4.1734.

12. Dhar M, Singh R, Manikonda SCK. From single chatbots to governed agent ecosystems: an agentic AI pattern catalogue and orchestration framework for mission-critical hospital information management systems. arXiv:2608.07627. 2026.

13. Cuneo J, Chun D, Khanna G. AI-GRACE: a use-case operationalization framework for agentic AI: from organizational objectives and obligations to deployment capabilities and architecture. arXiv:2609.21192. 2026.

14. Naik NK, Saroj AK, Poudel VP, Samantray S, Patel A. TRACCIA: an OpenTelemetry-based governance platform for AI systems. arXiv:2607.14309. 2026.