

Agentic AI for advanced therapies when every batch serves a unique patient

By Chris Wong, Raj Gopal Tripathi, Shubham Lathkar and Amol Agarwal



Impact where it matters.®

Agentic AI has crossed from slideware into live pharma operations. But the operating model that works for demand forecasting or procurement does not transfer cleanly to advanced therapies. In autologous cell therapy and radiopharmaceuticals, every batch or dose belongs to one named patient, no two treatment journeys are alike and a single scheduling decision carries clinical, manufacturing, quality, logistics and regulatory weight. That makes isolated optimization risky. A locally sensible decision can become wrong as quality, timing or logistics change.

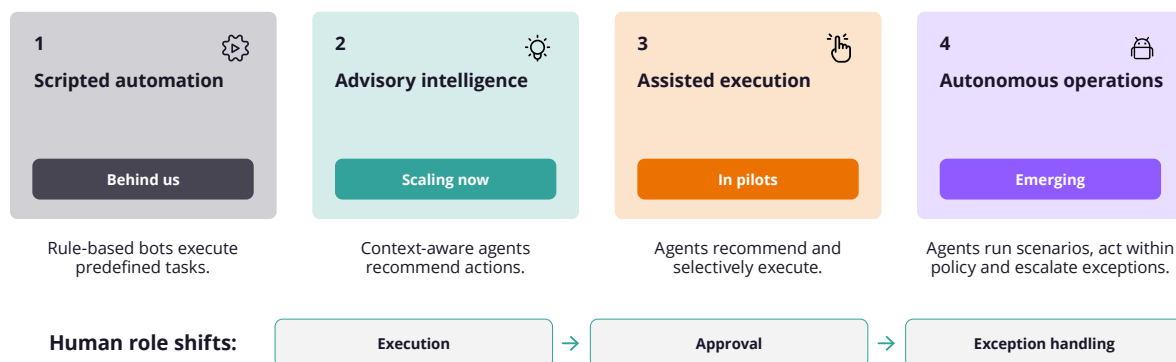
Advanced therapies therefore require a different agentic AI design principle. The coordination principle applies to both autologous cell therapy and radiopharmaceuticals, while the detailed operating examples below focus on autologous cell therapy. Specialist agents should not be built around isolated steps and connected later. They need to reason from a shared, patient-linked state, with orchestration determining when decisions can remain local and when they must be coordinated across the journey. Coordination first, automation second.

Agentic AI is moving from automation to controlled autonomy

Agentic AI is moving from scripted automation to advisory intelligence, assisted execution and, within defined policies, autonomous operations.

FIGURE 1:

Understand how human oversight changes as agentic AI gains autonomy





Agentic AI pursues outcomes and adapts as conditions change

An agentic system pursues a goal, perceives changing context, reasons over available information, acts through connected tools and escalates when the situation exceeds its authority.

When conditions change, an agentic system can reassess, recommend a different path, log its rationale and bring a human in with the right context.

Agentic AI works best in data-rich, logic-driven, traceable workflows

Pharma already offers proof points across demand forecasting, supplier intelligence, batch records, deviation triage and quality workflows, especially where work is data-rich, logic-driven, traceable and high-volume.

Cell therapy appears to fit that profile. The difference is that its decisions are patient-specific and tightly coupled, which makes step-by-step decomposition problematic.

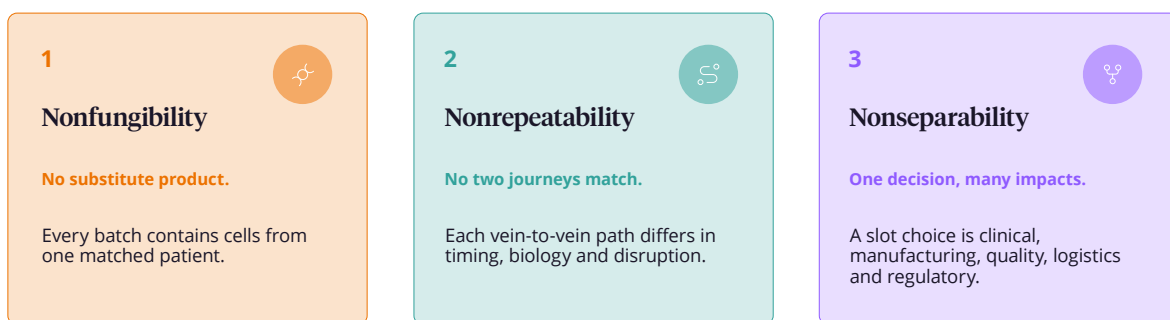
Patient-specific decisions complicate step-by-step agentic AI

At first glance, cell therapy seems easy to decompose into discrete agentic tasks: slot assignment, quality assessment, logistics coordination and chain-of-identity checks. But separating those decisions can obscure how changes in one step affect the others.

Three structural properties explain why.

FIGURE 2:

Understand why patient-specific therapy makes isolated agent decisions risky



First, nonfungibility. In a small-molecule supply chain, one lot can often substitute for another. In autologous cell therapy, there is no substitute. The batch is one named patient's own cells. A scheduling error or quality miss is not just an operational issue. It can directly affect whether that patient receives treatment within the available clinical window.

Second, nonrepeatability. No two vein-to-vein journeys are the same. Apheresis quality varies by patient. Manufacturing windows shift. Logistics disruptions can reshape the entire timeline. The process is not a repeatable sequence to optimize. It's a series of one-off patient journeys, each with its own constraints.

Third, nonseparability. A slot decision is never just a slot decision. It's also a clinical decision, a manufacturing decision, a quality decision, a logistics decision and a regulatory decision because patient readiness, capacity, starting material quality, cold-chain feasibility and chain of identity are all linked. These are not five sequential steps but five dimensions of the same decision.

Radiopharmaceuticals create the same coordination problem in a different operating context. The operating details differ, but each patient-linked dose still requires scheduling, manufacturing, quality, logistics and regulatory constraints to stay aligned as conditions change.

Together, nonfungibility, nonrepeatability and nonseparability change the design problem. The objective is not to optimize a series of independent tasks. It's to preserve one patient-linked decision context as constraints change. An isolated agent can make a locally rational choice that becomes wrong once quality, timing or logistics shift. In cell therapy, coordination failure is the failure mode.

Consider a single manufacturing batch failure on day six of a 14-day process. The obvious question is whether to remanufacture, but the real decision spans the whole patient-linked system. Can remanufacturing happen quickly enough for the product to be ready by the original promised date? Can the schedule be restructured without unacceptable knock-on effects for other patients? Is there another site, a cancellation or a buffer to draw on, and can logistics claw back the lost time? Each answer changes the others, and all of them have to be weighed while the window for action is still measured in hours.

This type of disruption illustrates why the design principle has to be coordination first, automation second.

A shared patient-linked state keeps specialist agents coordinated

The solution is not more agents but an altogether different architecture. A central orchestrator agent maintains one continuously updated, patient-linked state across the vein-to-vein journey: patient status, manufacturing slots, quality signals, batch disposition, cold-chain status and bedside readiness.

The orchestrator does not replace specialist agents. It coordinates them. When a disruption occurs, every relevant agent reasons from the same live state rather than from separate handoffs. Slot management sees quality flags, quality assessment sees slot constraints, logistics sees disposition status and patient orchestration sees the revised treatment timeline.

Importantly, coordination is not the same as constant collaboration. Plenty of routine work (a status confirmation, a checklist, a shipment acknowledgment, an audit-log update or a standard document) never needs systemwide reasoning. Orchestration earns its keep only when a decision starts to touch several dimensions of the patient journey at once. The goal isn't to orchestrate everything, only the decisions that matter.



5 specialist agents operate around one patient-linked reality

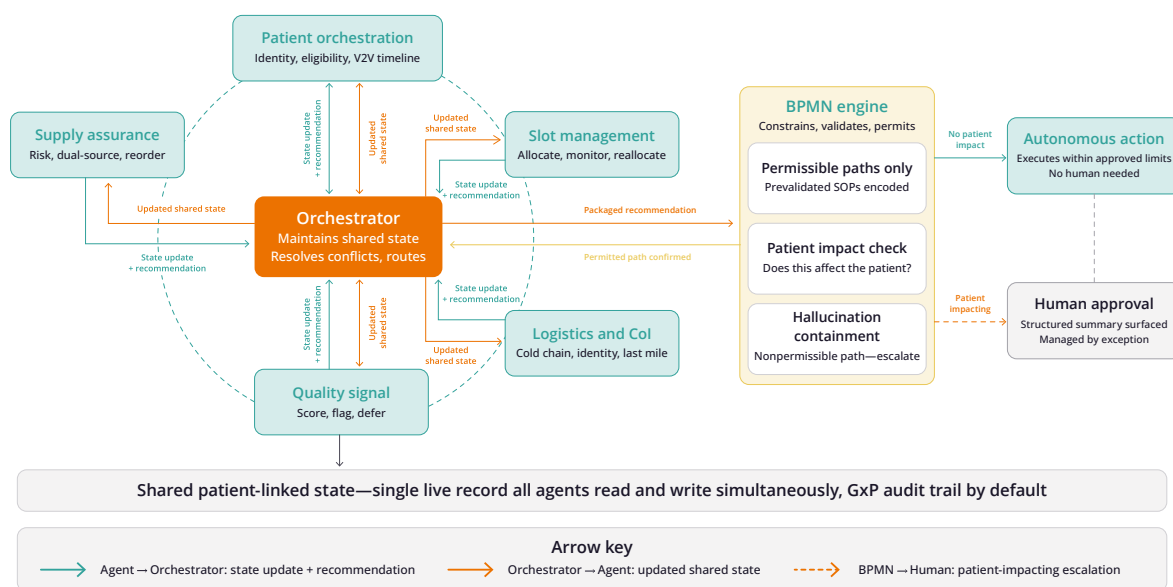
Five purpose-built specialist agents operate around the same patient-linked state.

- **Patient orchestration agent:** Maintains the live patient-linked record, including eligibility, treatment center readiness, conditioning windows and the evolving vein-to-vein timeline so every agent reasons from the latest patient context.
- **Slot management agent:** Allocates and reallocates manufacturing capacity across sites and contract development and manufacturing organizations using patient timing, quality signals, network constraints and disruption scenarios, turning manual firefights into fast, evidence-backed options.
- **Quality signal agent:** Connects apheresis and in-process quality data directly to manufacturing decisions. It flags borderline collections, transit excursions and fit-for-manufacturing risks early enough to influence slot decisions before capacity or critical materials are consumed.

- **Supply assurance agent:** Monitors upstream risk across vectors, plasmids, cytokines, media and closed-system consumables. It treats supply security as a patient-access issue, surfacing supplier constraints and mitigation options before they threaten active or near-term patient journeys.
- **Logistics and chain of identity agent:** Coordinates apheresis and finished goods movement while verifying chain of identity (Col) at each custody handoff. It aligns shipment timing with patient readiness, monitors exceptions in real time and ensures the Col record is complete and inspection-ready by default.

FIGURE 3:

Understand how the shared state keeps agent decisions coordinated and patient-safe



Instead of coordination traveling by report, email or delayed handoff, each specialist agent reads from and writes to the same patient-linked state, while the orchestrator coordinates decisions across it.



Shared state, guardrails and human judgment keep the network safe

Building five well-designed specialist agents is only half the problem. The harder question is how to make them work together reliably in a good practice (GxP)-regulated environment where every decision can have patient, quality, logistics and regulatory consequences. The network must solve three problems at once: coordination, containment and escalation.

Coordination happens through shared state, with the orchestrator resolving conflicts by defined priorities and keeping every agent on one source of truth.

Containment happens through guardrails. An agent can produce an intelligent-looking recommendation that breaks a timeline commitment, an approved site network, a validated shipping lane or a documented standard operating procedure. In a regulated environment, the guardrail cannot sit after the decision. It has to be part of the architecture. This is where a business process model and notation (BPMN) engine is essential. The agents reason, evaluate options and recommend, while the BPMN layer decides which recommendations are permissible, which can execute automatically and which must escalate. In a batch failure, for instance, the slot management agent might weigh several recovery paths, but the BPMN engine confirms the chosen one maps to a preapproved response category before anything executes.

Escalation happens by exception, based on patient impact rather than decision type. Any decision that materially affects a patient's treatment window, product integrity or safety needs human approval. Decisions that simply optimize within the validated plan, such as swapping a carrier within an approved lane or adjusting a noncritical buffer, can proceed automatically. This distinction creates real oversight rather than the appearance of it, because routing low-stakes decisions through approval adds delay without adding protection and creates the fatigue that leads to reflexive sign-offs.

When a human is needed, the escalation should arrive already prepared with what happened, the available options, the recommended action, the rationale, the downstream impact and the time left before choices narrow. The approver can then review a structured recommendation and either authorize it or request a rerun under a different constraint. That is what “managed by exception” means. Human authority does not diminish. It becomes more focused.

Beneath this sits the audit trail, which must be designed from the start. Every recommendation, BPMN path selection, escalation trigger, autonomous action and human approval should be captured with timestamp, rationale and source. In a GxP environment, this record is not paperwork. It is the evidence that the process stayed in control.

A safe network is therefore not just an AI build. It's an operating-model build that brings process owners, quality, supply chain and technology teams together. The agents provide the reasoning, while shared state, guardrails, escalation logic and auditability make the network safe enough to use.

Which agentic AI activities need human oversight in advanced therapies?

Three areas should remain with people as agentic AI scales in advanced therapies: patient-prioritization decisions, consequential quality and regulatory actions, and novel situations outside approved operating logic.

Prioritizing between equally urgent patients competing for one slot is an ethical judgment, not a scoring exercise. Batch disposition requires quality-unit review and approval before release, while biologics license applications require an authorized signatory. Deviation closure and other regulatory responses do not follow one universal sign-off rule, so they should remain within applicable quality-system and submission approval pathways.

Novel situations that do not match an encoded BPMN path, such as a new supply shock or an unprecedented failure mode, also require human judgment. The network can rank options, assemble evidence and surface what matters, but a person makes the final call.

Agent-ready data must be current, patient-linked and shared

Every agentic AI conversation eventually reaches the same hard truth. The data is not ready. For cell therapy, that concern is usually valid because stale, incomplete or disconnected signals produce recommendations that sound coherent but can be operationally wrong.

A functioning network needs five data streams that are real-time, patient-linked and accessible without manual intervention: patient status, apheresis quality, manufacturing progress, upstream supply risk and cold-chain and chain-of-identity telemetry. Most of these exist in some form today, but rarely in the integrated form that reliable reasoning requires.

Start with advisory agents and expand autonomy as evidence grows

The leadership implication is architectural. Don't deploy specialist agents function-by-function and assume orchestration can be added later. Define the patient-linked state, decision rights, BPMN guardrails and escalation logic early. Then start narrow, prove value and expand autonomy as data quality and performance evidence mature.

Stage 1: Assisted—agents that see and surface. Agents operate in an advisory mode.

They monitor available patient, quality, manufacturing and logistics data, surface risks, generate options and package recommendations for human approval. The focus is visibility, decision support and evidence generation, not autonomous execution.

Stage 2: Augmented—agents that act within boundaries. Validated paths begin to execute automatically within BPMN guardrails. Human approval remains for patient-impacting decisions, low-confidence recommendations and exceptions outside the approved operating logic. The focus shifts from recommendation quality to controlled execution and auditability.

Stage 3: Autonomous within guardrails—agents that manage by exception.

Routine decisions flow automatically across the network. Humans focus on novel, consequential or ambiguous decisions. Ongoing monitoring ensures BPMN paths, knowledge rules and autonomous actions remain validated as products, sites and geographies evolve.

Clinical and commercial teams converge on the same coordinated architecture

While the maturity path is common, organizations don't start from the same place.

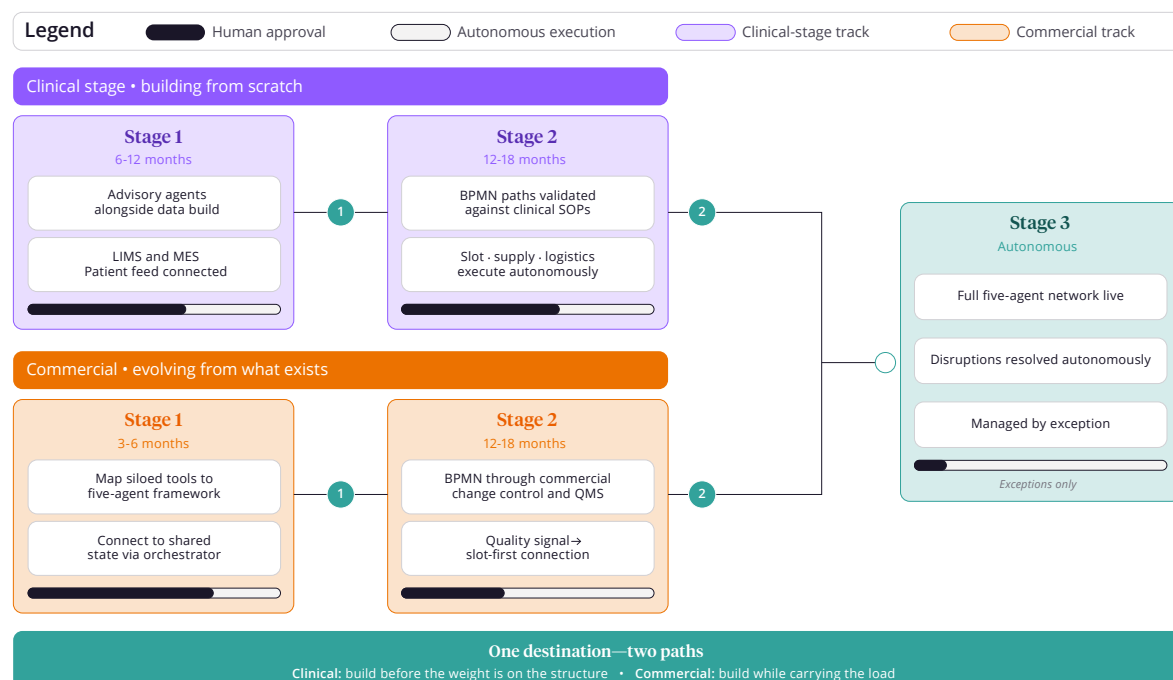
For clinical-stage sponsors, the opportunity is architectural. Data models, integrations, operating procedures and governance frameworks are still being shaped. Decisions made during Phase 2 and Phase 3 determine how easily the organization will scale once commercial demand arrives.

Commercial-stage operators face a different challenge. Their supply chain is already live, patients are being treated and critical processes are already in motion. Transformation must happen around an operating system rather than in place of it. Often, organizations already have pieces of the future state in place, but those capabilities exist as disconnected tools, workflows and data streams rather than as a coordinated network.

The advantage of the clinical-stage sponsor is flexibility. The advantage of the commercial operator is operational experience. Despite these different starting points, both paths converge on a coordinated, BPMN-governed, patient-linked network that manages routine decisions autonomously and escalates only those decisions that materially affect the patient journey.

FIGURE 4:

Compare how clinical and commercial teams can scale autonomy safely



Build the coordination layer before commercial scale

For Phase 2 and Phase 3 sponsors, the roadmap is not a future commercial topic. The decisions being made today, from system integrations and patient data models to escalation logic and operating procedures, will determine what is possible at launch.

Sponsors who begin with advisory agents during clinical development can arrive at launch with integrated data, validated BPMN paths and an organization that already understands how to govern an agentic operating model. Those who delay risk scaling a supply chain that was designed for dozens of patients into one that must support hundreds.

Build for the patient journey first and automate the process second

For advanced therapies, scaling autonomy starts with an operating model that keeps decisions connected to the patient. Patient-linked data, orchestration logic, BPMN guardrails and clear human decision rights provide the structure agents need to act safely as conditions change.

Clinical-stage sponsors can build that structure while processes and systems are still taking shape. Commercial operators can connect capabilities that already exist but remain fragmented across functions. In either case, the objective is the same: Give specialist agents a common context for making and escalating decisions.

The organizations that get this right will not simply automate more work. They will be better equipped to respond when quality, capacity, logistics or patient readiness changes—and to do so without losing sight of the patient whose treatment depends on the decision.



About ZS

ZS is a management consulting and technology firm that partners with companies to improve life and how we live it. We transform ideas into impact by bringing together data, science, technology and human ingenuity to deliver better outcomes for all. Founded in 1983, ZS has more than 15,000 employees in over 40 offices worldwide.

Learn more: visit www.zs.com or follow us on [LinkedIn](#).

