



INTUITIONLABS / RESEARCH & PRACTICAL GUIDANCE

Claude AI Pharma R&D Automation Index: Autonomy and Oversight

Build practical AI for pharma and biotech with IntuitionLabs. We help life-science teams turn complex information and workflows into useful software, governed knowledge systems and AI tools.

Published by IntuitionLabs · 2026-10-27 · [claude ai pharma r&d automation](#) · [ai agent autonomy in pharma r&d](#) · [pharma r&d automation index](#) · [ai agent governance metrics](#) · [human oversight of ai agents](#) · [anthropic r&d automation index](#) · [claude for life sciences](#) · [ai agent safety in pharmaceutical research](#) · [measuring ai agent autonomy](#) · [fda ai credibility framework](#)

Original article: <https://intuitionlabs.ai/articles/claude-rd-automation-index-pharma-agent-autonomy>



In this article

1. Executive Summary
 2. Introduction and Background
 3. What Anthropic Published on September 17, 2026
 4. What Not to Infer
 5. Translating AL0 to AL5 into Pharma R&D
 6. Building the Pharma R&D Automation Index
 7. Oversight Metrics, Assurance Resources and Governance Triggers
 8. Credibility Versus Autonomy: Keeping Two Questions Apart
 9. Data Analysis and Evidence
 10. Implications and Future Directions
 11. Frequently Asked Questions (FAQs)
 12. Conclusion
-

Executive Summary

On September 17, 2026, Anthropic published a prototype **Anthropic R&D Automation Index** that rates how much of its own AI research and development (R&D) work Claude performs (www.engadget.com). As of August 2026, Claude "leads" **26%** of that work, the share at or above "collaborates" exceeds **90%**, and no measured subset is fully autonomous (apnews.com) (www.anthropic.com). The oversight data cover one internal platform with about **30,000** agents at any one time (www.anthropic.com).

These figures describe an AI laboratory measuring itself. They are not a pharma benchmark, productivity lift, error rate or compliance result. The judge agreed exactly with staff ratings **59%** of the time, against **35%** between humans, and fell within one level **97%** of the time, but those numbers characterize Anthropic's exercise only (www.anthropic.com). Anthropic plans to embed third-party evaluators (www.anthropic.com). No independent replication was found as of October 27, 2026.

For pharma R&D, the method transfers better than the numbers. This report proposes a **Pharma R&D Automation Index**: a frozen, versioned task basket; autonomy ratings tied to evidence of who scoped, handled exceptions, verified, approved and deployed; weights based on person-time, consequence or spend; and a raw distribution published beside any single score. Oversight (coverage, review latency, escalation, override, rollback) is tracked separately, because Anthropic reports monitor coverage (every action on one platform passes an online monitor) and a 0.002% block rate as measures distinct from automation level (www.anthropic.com).

Autonomy must also be kept apart from credibility. The U.S. Food and Drug Administration (FDA) draft guidance on AI for regulatory decision-making asks whether a model is credible for a defined context of use, scored by model influence and decision consequence. The guidance is a draft distributed for comment (www.fda.gov). The same draft excludes AI used in drug discovery and some operational uses from its scope (www.fda.gov). Surveys show uneven adoption: 64% of 200 surveyed life-sciences executives use AI regularly

in R&D, and 75% was the Phase I success rate for AI-enabled emerging-biopharma programs in IQVIA's 2026 analysis (www.whitecase.com) (www.iqvia.com). Governance triggers (testing windows, independent review, narrower permissions, rollback, pause) should key on autonomy, consequence and drift together.

Introduction and Background

Pharma R&D leaders are being asked a question that existing validation vocabulary only partly answers: how much of the work is now done by AI agents, and who is still deciding? Classical computerized-system validation asks whether a tool performs as intended. Model credibility asks whether an output can be trusted for a stated purpose. Neither measures the shift from a person using an assistant to an agent that plans, executes, handles surprises and hands back a finished package for sign-off.

On September 17, 2026, Anthropic published a measurement approach for exactly that shift inside its own laboratory (www.engadget.com). It reports three things: how much AI R&D is performed by Claude, how well agent actions are overseen, and how compute is allocated between safety and other work. Anthropic describes these as measurements of how models are built. Major wire coverage appeared within a day, including an Associated Press (AP) report on September 18 (apnews.com). Anthropic frames the measurements around recursive self-improvement, the idea of a model building its own successor. That framing leaves a gap for regulated industries, which need to know how to adapt the measurement logic to scientific and regulatory work without borrowing the headline numbers.

This report is written by IntuitionLabs, a life-sciences and AI consultancy. IntuitionLabs is an adjacent advisor, and this report does not offer an index product or agent software. The analysis is a framework proposal, and nothing here is endorsed by Anthropic, FDA, the National Institute of Standards and Technology (NIST) or any other body named. Readers who need legal, quality or regulatory determinations should consult qualified professionals in those fields.

The report first states what Anthropic measured, with dates and scope, then what not to infer. It translates the six-level scale into pharma examples, builds the index mechanics, specifies oversight metrics and governance triggers, and compares the proposal with regulatory sources and pharma adoption evidence. The target readers are heads of data and AI, R&D and translational leaders, clinical and regulatory operations, AI governance councils, internal audit, quality and model-risk teams.

What Anthropic Published on September 17, 2026

The scale and the headline figures

Anthropic uses an Automation Level (AL) rating scale developed by Epoch AI (www.engadget.com). The scale runs from AL0 (no AI involvement) to AL5 (AI operates fully autonomously, with no human in the loop), with intermediate levels named minimal AI involvement, AI assists, collaborates, leads and autonomous (www.anthropic.com). At AL4 ("leads") AI can complete most of the task end to end from a high-level prompt while the human supervises (www.anthropic.com). Epoch's own write-up rates tasks on a 0 to 5 scale and calls the ratings subjective (epochai.substack.com).

Anthropic's footnote illustrates the distinction with a broken nightly data pipeline. At AL4 the engineer hands over the failure alert, Claude fixes and tests the pipeline, and the engineer then reads the write-up and decides whether the fix ships (www.anthropic.com). AL5, in which the issue is detected, fixed and deployed with no human involvement, is described as a level not yet reached. AP summarized the same point as the model not yet working completely autonomously (apnews.com).

The three August 2026 findings are: Claude is not fully autonomous on any measured subset of work, Claude leads 26%, and more than 90% sits at collaborates or above (apnews.com) (www.anthropic.com). Anthropic also urged other developers to publish comparable metrics regularly (apnews.com).

How the index was built

The index rests on a task map, a rating method and a weighting method. Anthropic built the map bottom-up: for each week in July 2026 it randomly sampled 20% of staff in each department in the model R&D loop, and a Claude research agent listed the tasks each person worked on (www.anthropic.com). The resulting tree has 542 nodes, of which 378 are leaves, and it is frozen so that every measurement uses the same basket of work (www.anthropic.com). For weighting, each person receives one unit of weight per week, split evenly across the tasks worked on that week (www.anthropic.com).

Oversight and compute snapshots

On oversight, Anthropic defines three metrics: coverage (the share of actions passing through a monitor), review latency, and escalation rate. As of August 2026 the measurements covered about 30,000 agents on the most-used internal platform, and only that platform. All actions pass through an online monitor before execution (www.anthropic.com). Of more than a billion decisions in August 2026, 0.002% (about 1 in 47,000) were blocked (www.anthropic.com).

On compute, about 6% of AI R&D compute and about 12% of compute for AI-driven AI R&D went to safety in the week of July 13 to July 20 (www.anthropic.com).

What Not to Infer

Several inferences are tempting and unsupported.

- **Not a pharma benchmark.** The 26% describes Anthropic's model-development work, and its task tree is built from AI-laboratory work, not drug R&D. Pharma shares at "leads" will depend on each organization's tasks, systems and review culture.
- **Not a productivity lift.** The index rates who does a task, not how well, how fast or at what cost.
- **Not an error rate.** The 0.002% block rate describes one monitor on one platform, not agent accuracy.
- **Not a compliance result.** Nothing in the post addresses GxP (good practice regulations) or any regulator's expectations.
- **Not independent validation.** The index figures are Anthropic-reported, and no independent replication was found as of October 27, 2026. Anthropic acknowledges that using its own models to evaluate its systems risks shared errors (www.theneuron.ai). Engadget reports that Anthropic believes other labs could reproduce the process with their own data and third-party validation (www.engadget.com).

Judge figures describe one exercise. The judge agreed exactly with staff 59% of the time versus 35% between humans, and 97% of ratings fell within one level (www.anthropic.com). Those are properties of Anthropic's judge, task tree and staff, not general model accuracy.

Verification status as of October 27, 2026. Searches for this report found no independent replication of the 26%, the oversight figures or the compute shares. Anthropic says it plans to embed independent evaluators with access to internal processes and data, which is a plan rather than a result (www.anthropic.com). A negative search result is not proof that no verification exists, so readers should re-check. The post's page text carries no visible date, so the September 17 date rests on Engadget's timestamp and AP's report of a Thursday announcement (www.engadget.com) (apnews.com).

Translating AL0 to AL5 into Pharma R&D

The six levels transfer if each is defined by who does what at each stage of a workflow. Five roles matter: who scopes the work, who handles exceptions, who verifies results, who approves action, and who deploys or submits. Separate research supports a role-based view: Feng, McDonald and Zhang define five user roles (operator, collaborator, consultant, approver, observer) and argue that autonomy is a deliberate design decision separate from capability (arxiv.org); at their top level the user's only control is an emergency off-switch (arxiv.org). The Organisation for Economic Co-operation and Development (OECD) similarly defines autonomy as the degree to which a system can learn or act without human involvement (oecd.ai), and NIST's framework describes AI systems as operating with varying levels of autonomy (nvlpubs.nist.gov). The driving-automation standard SAE J3016 offers a familiar analogy, with a six-level taxonomy from no automation to full automation (webstore.ansi.org).

Table 1 below adapts the scale using literature screening as a running example. The examples and prohibited actions are this report's illustrative design, not requirements from Anthropic, Epoch AI or any regulator.

Level	Pharma example (literature screening)	Required human role	Prohibited at this level (illustrative)
AL0 No AI involvement	Scientist screens abstracts manually.	Performs all steps.	Not applicable.
AL1 Minimal	Agent suggests search terms; scientist runs the searches.	Performs all steps.	Agent access to controlled repositories.
AL2 Assists	Agent summarizes papers on request; scientist drives and reviews everything.	Scopes, executes, reviews every output.	Agent-initiated actions; writes to shared systems.
AL3 Collaborates	Agent screens a set under close direction; scientist reviews line by line.	Scopes, directs, resolves exceptions, verifies, approves.	Agent decisions on exceptions, release or submission.
AL4 Leads	Agent screens, ranks and writes up findings from a high-level request; scientist reads the write-up and decides.	Defines the request, reviews write-up and evidence, approves any action.	Deployment, submission or changes to controlled records without a named approver.

Level	Pharma example (literature screening)	Required human role	Prohibited at this level (illustrative)
AL5 Autonomous	Agent monitors new literature, scopes, executes and acts with no human involvement.	None required; oversight by monitoring and an emergency stop.	Not permitted for Tier 3 consequence (patient, regulatory or release impact) without documented risk acceptance and independent review.

The table shows why the AL3 to AL4 boundary carries the most weight. At AL3 a person stays tuned in and resolves surprises. At AL4 the agent handles surprises and documents them, and a person decides at the end. Because exact agreement among trained raters was only 35% in Anthropic's exercise, each pharma rating needs written evidence, not impressions.

The same scale applies to other workflows, with different risk profiles:

- **Assay and bioinformatics pipeline changes.** The nightly-pipeline example maps almost directly; risk depends on whether the pipeline feeds regulated data.
- **Candidate prioritization.** An agent can rank candidates at AL4, but rankings influence portfolio decisions, so bias in the evidence set matters.
- **Protocol and analysis-plan drafting.** Anthropic lists drafting study protocols and standard operating procedures (SOPs) as a use case with its Benchling connector (www.anthropic.com). Draft stage is not approval stage.
- **Regulatory content.** Novo Nordisk's published case study states a goal of automating the entire process of producing Common Technical Documents, a goal and not an achieved result (claude.com).
- **Release or submission decisions.** These are approval and deployment actions and should stay human-approved unless a documented, independently reviewed risk acceptance exists.

Building the Pharma R&D Automation Index

Step 1: Build a bottom-up task inventory

Anthropic derived tasks from work records and sampled staff each week, not from manager lists. The pharma equivalent draws on SOPs, ticketing systems, electronic lab notebooks, study plans and work records. Privacy needs explicit handling: sample at the level of task descriptions, strip personal data, keep patient-level and commercially restricted content away from rating agents, and obtain privacy review where applicable. A usable hierarchy is **function** (translational research) → **process** (target assessment) → **task** (rank candidate list) → **evidence** (tickets, logs, notebook entries) → **owner** → **autonomy level** → **consequence tier**.

Step 2: Freeze, version and define materiality

Anthropic freezes its tree so measurements are comparable over time. A pharma program should publish a basket version (for example, v2026.4), change it only through change control, and report old and new baskets in parallel for one cycle after a rebuild. A suggested consequence scheme:

- **Tier 1:** low consequence, internal and reversible (literature triage, internal summaries).
- **Tier 2:** affects study design, analysis or GxP-relevant records, but is reviewed before use.

- **Tier 3:** affects patient safety, product quality, regulatory submission content or release decisions.

The scheme is a proposal. It parallels, but does not replace, the decision-consequence concept in FDA's draft guidance, which concerns the significance of an adverse outcome from an incorrect decision (www.fda.gov).

Step 3: Rate autonomy from evidence

Anthropic has a research agent gather evidence on who does each kind of work, and an independent Claude judge then assigns one of six levels (www.anthropic.com). A pharma program can follow that structure with a person in the loop. For each task node, require evidence on five questions: who scoped the work, who handled exceptions, who verified results, who approved the action, and who deployed or submitted. Rate the node at the **lowest** level that those answers support. If an agent does everything but a named person approves every deployment, AL4 is supported; AL5 requires that the agent also deploys, with monitoring and stop controls in place. Each rating carries an evidence reference and a rater identity (human or model) so disagreements can be audited.

Step 4: Choose weights and understand their biases

Anthropic used person-time as a proxy for importance. Pharma programs have at least four options:

- **Person-time.** Easy to explain. Bias: overweights labor-intensive, low-consequence work and underweights short high-stakes decisions such as a release approval.
- **Business criticality.** Weights by program value or timeline impact. Bias: reflects management judgment and may lag reality.
- **Patient or regulatory consequence.** Weights by consequence tier. Bias: shifts the index toward work least likely to be automated, which suits governance but understates adoption.
- **Spend or compute.** Weights by cost. Bias: a compute share measures what is spent, not effort, so efficiency gains can move the number without changing the work.

Publish at least two weightings. A gap between person-time and consequence-weighted scores is itself a finding.

The calculation and the raw distribution

The weighted index is $\text{Index} = \frac{\text{sum}(\text{task weight} \times \text{normalized autonomy level})}{\text{sum}(\text{task weights})}$, where the normalized level is AL divided by 5. Anthropic's headline figures are shares of work at given levels, not a mean score. A mean can hide a distribution, so publish the raw weight share at each level alongside it.

(Hypothetical Example) Take six task families with weights summing to 1.0: literature screening (0.20, AL4), pipeline changes (0.25, AL3), candidate prioritization (0.15, AL2), protocol drafting (0.15, AL3), regulatory content (0.15, AL2) and release or submission decisions (0.10, AL1). Normalized levels are 0.8, 0.6, 0.4, 0.6, 0.4 and 0.2, and the weighted sum is $0.16 + 0.15 + 0.06 + 0.09 + 0.06 + 0.02 = 0.54$. The distribution is more informative: 60% of weight sits at AL3 or above, 20% at AL4 or above, none at AL5, and the highest-consequence family sits at AL1. The figures are invented to demonstrate arithmetic.

Confidence and rating quality

An index without confidence information invites over-reading. *Table 2* below lists the rating-quality measures to publish with each release, with Anthropic's figures shown only as reference points for what its exercise reported.

Measure	Definition	Anthropic reference point (not a target)	Why it matters
Sample size	Sampled tasks and staff per function	20% of staff per department per week, July 2026	Coverage of the real task basket
Exact agreement, model vs human	Share of nodes where judge and expert give the same level	59%	Judge reliability
Exact agreement, human vs human	Same, between two expert raters	35%	Ceiling on what the judge can match
Within-one-level agreement	Share of pairs differing by at most one level	97%	Gross misrating risk
Missing evidence	Share of nodes rated without full five-question evidence	Not reported	Rating validity
Classifier or human disagreement	Nodes where model and expert differ by two or more levels	Not reported	Triggers re-review

Anthropic's numbers come from one organization, one judge model and one task tree. A pharma program should measure its own values and set its own acceptance limits before using the index for decisions.

Oversight Metrics, Assurance Resources and Governance Triggers

The oversight dashboard

Oversight should be reported separately from automation, as Anthropic does by defining coverage, latency and escalation as distinct measurements. Suggested pharma metrics:

- **Action coverage** = monitored actions / total actions. Anthropic reports that all actions on one platform pass through an online monitor before execution.
- **Pre-action versus post-action review.** Record which monitors act before execution and which afterward, and which harms each is meant to catch.
- **Review latency.** Report the median and the 95th percentile (p95) time from action to review.
- **Escalation rate** = flagged or blocked actions / monitored actions, with blocked and redirected actions counted separately. Anthropic's blocked share is 0.002%.
- **Human escalation.** Count items reaching a person.
- **Human override rate** = outputs changed or rejected by reviewers / outputs reviewed.
- **Rollback rate** = actions reversed / actions deployed.

Override data need careful reading. In clinical decision support, a systematic review found average alert override rates between 46.2% and 96.2%, so a high rate can reflect alert fatigue rather than reviewer diligence (www.ncbi.nlm.nih.gov). Automation bias, the tendency to over-rely on automation, is a further known risk, examined in a systematic review that screened 13,821 papers and included 74 (pmc.ncbi.nlm.nih.gov). The EU AI Act's human-oversight article lists awareness of automation bias and the ability to disregard or reverse output among the capacities overseers must have (artificialintelligenceact.eu). A near-zero override rate can therefore signal excellent output or rubber-stamping, so quality teams should sample reviewed items independently.

Anthropic's separate research on agent use shows review behavior shifting with experience: roughly 20% of sessions among new Claude Code users use full auto-approve (www.anthropic.com). Capability is also moving: METR's analysis found the 50% time horizon of frontier models doubling approximately every seven months since 2019 (arxiv.org), which argues for rebuilding a frozen basket periodically. These are product observations, not pharma data.

Assurance-resource share

Anthropic's third measurement asks what fraction of compute goes to safety work. A pharma analogue is **Assurance-resource share = compute and labor tagged to validation or safety / total agent-program resources**, with compute and labor reported separately. Anthropic's snapshot of about 6% (about 12% for AI-driven work) shows the form of the output, not a target.

The boundary is ambiguous. Anthropic says its safety definition is one reasonable choice among many and that some cases stayed hard to classify even after hours of human review (www.anthropic.com). Pharma equivalents include a validation harness that also speeds throughput, or an evaluation set used for both release testing and product improvement. Define category rules in writing, publish them with the number, and have quality or internal audit sample the tagging.

Governance triggers

Anthropic suggests that such measures could trigger stronger requirements, such as a fixed testing window before a new model is used for further AI R&D (www.anthropic.com). A pharma translation combines autonomy level, consequence tier and performance drift. *Table 3* below is an illustrative trigger matrix; thresholds are placeholders for each organization to calibrate and are not benchmarks from any source.

Autonomy level	Consequence tier	Condition	Required control
AL0 to AL2	Any	Stable	Standard periodic review.
AL3	Tier 1	Stable	Quarterly sampled review of outputs.
AL3 or AL4	Tier 2	Level rises within a quarter, or a new model or tool version is deployed	Enhanced testing window before the higher level takes effect, plus independent review of ratings.
AL4	Tier 3	Any	Named human approval before any action with external effect; narrower permissions (read-only by default).

Autonomy level	Consequence tier	Condition	Required control
AL5 observed	Tier 2 or 3	Any	Temporary pause of expansion pending independent review and documented risk acceptance.
Any	Any	Override, escalation or sampled error rate leaves its predefined band for two consecutive periods	Investigate; narrow permissions or roll back to the prior validated configuration.
Any	Tier 3	Action coverage below the predefined minimum, or p95 review latency above limit	Pause autonomous actions in that scope until restored.

Level alone is not a trigger: AL4 on literature triage and AL4 on a submission step need different controls. Drift can also justify tighter control without any change in level. Permission design can implement narrower scope. Anthropic's Claude Code documentation describes permission modes that set which actions run without asking first, including a mode where a second model reviews actions instead of the user and a bypass mode listed for isolated containers and virtual machines only (code.claude.com). Its Claude Science announcement describes an agent that asks before reaching new resources and lets users review or revoke decisions (www.anthropic.com). These are vendor descriptions of product design, offered as examples of gradable permissions.

Quarterly audit and change-control cycle

In the first month, refresh the sampled inventory and compare it with the frozen basket to detect new kinds of work. In the second month, re-rate nodes with written evidence and compute agreement statistics. In the third month, compute the index under at least two weightings, publish the raw distribution and oversight dashboard, and review trigger conditions. Any change to the basket, rubric, judge model or weights goes through change control with a version bump. Internal audit should re-perform a sample of ratings and confirm that dashboard data come from system logs, not self-reporting.

Credibility Versus Autonomy: Keeping Two Questions Apart

An agent's autonomy level answers "who does the work and who decides?" A model's credibility answers "can this output be trusted for this stated purpose?" The two are related and not interchangeable.

FDA's draft guidance on AI to support regulatory decision-making offers a risk-based credibility assessment framework for a particular context of use (www.fda.gov). It sets out a seven-step process that begins with defining the question of interest and the context of use (www.fda.gov). Model risk combines model influence and decision consequence (www.fda.gov). The document carries a January 2025 date and a "Not for Implementation" header (www.fda.gov), says it is distributed for comment purposes only, and FDA's drug-AI page, accessed October 27, 2026, still refers to a draft guidance published in 2025 (www.fda.gov). It should be treated as draft and not for implementation. No source reviewed indicated finalization, but readers should check FDA's current guidance listing.

Scope matters. The draft does not address AI models used in drug discovery, or for operational efficiencies that do not affect patient safety, drug quality or the reliability of study results (www.fda.gov). An agent drafting content may therefore fall outside that guidance and still need controls under a company's quality system. FDA says its experience with over 500 submissions with AI components from 2016 to 2023 informed the draft (www.fda.gov).

The two questions combine as follows. A tightly specified model with strong credibility evidence may sit at AL2, while a loosely specified general-purpose agent may reach AL4 on low-consequence work without any credibility file. A workflow with high model influence and high decision consequence needs strong credibility evidence at any autonomy level, and the same workflow rated AL4 needs added controls over exceptions, approvals and deployment. Neither FDA nor NIST has endorsed Anthropic's index, and no source reviewed suggests otherwise.

Related frameworks and status

- **FDA and EMA principles.** FDA lists 10 principles tailored to the drug development cycle, including human-centric design (www.fda.gov) and a clear context of use (www.fda.gov), developed with the European Medicines Agency (EMA) (www.fda.gov).
- **EMA reflection paper.** Adopted in final form by the Committee for Medicinal Products for Human Use on 9 September 2024, it focuses only on AI in the medicinal product lifecycle (www.ema.europa.eu), recommends a risk-based approach to development, deployment and performance monitoring, and says a human-centric approach should guide all development and deployment (www.ema.europa.eu). EMA's network workplan for 2025 to 2028 includes guidance on AI throughout the medicine lifecycle (www.ema.europa.eu).
- **EU GMP draft Annex 22.** The consultation closed on 7 October 2025 and the draft addresses manufacturing, not R&D (health.ec.europa.eu). It covers static models only, states that it does not apply to generative AI and large language models, and keeps qualified personnel accountable for outputs in non-critical uses (picscheme.org). It also calls for change control and monitoring for drift in input data (picscheme.org), and its sponsors describe it as foreseeing continuous oversight of AI systems (picscheme.org).
- **21 CFR Part 11 and GAMP.** Electronic-record rules require validation and secure, time-stamped audit trails (www.ecfr.gov), which agent action logs should respect. The International Society for Pharmaceutical Engineering describes its GAMP Guide: Artificial Intelligence as a holistic interpretation for AI-enabled computerized systems that focuses on risk-based effort (ispe.org).
- **ICH E6(R3).** The International Council for Harmonisation guideline encourages a risk-based, proportionate approach and states that computerised systems used in clinical trials should be fit for purpose (database.ich.org); the ICH document history records Annex 2 endorsement under Step 4 on 3 June 2026 and the consolidated E6(R3) text, including Annex 2, as adopted on 16 June 2026 (database.ich.org).
- **NIST.** The AI Risk Management Framework is voluntary, defines four functions and, per NIST's page, is being revised; its generative AI profile was released on July 26, 2024 (www.nist.gov) (nvlpubs.nist.gov) (www.nist.gov). In February 2026 NIST announced an AI Agent Standards Initiative focused on agents capable of autonomous actions (www.nist.gov).

- **EU AI Act.** Article 14 says human oversight shall aim to prevent or minimise risks to health, safety or fundamental rights (artificialintelligenceact.eu) and requires oversight measures commensurate with the risks, level of autonomy and context of use for high-risk systems (artificialintelligenceact.eu), and the Commission lists logging for traceability among high-risk obligations (digital-strategy.ec.europa.eu). The Commission says the Act became applicable on 2 August 2026, with an extended transition for systems embedded in regulated products until 2 August 2028 (digital-strategy.ec.europa.eu). The Commission page also states that Annex III high-risk use cases have been extended to 2 December 2027 under the AI Omnibus Regulation, which entered into force on 27 July 2026 (digital-strategy.ec.europa.eu), so readers should confirm scope and dates against the final text of the Regulation. Whether a given R&D workflow is high-risk is a legal classification question for counsel.

Data Analysis and Evidence

Public data on how autonomously agents work in pharma R&D are thin. Context figures show why consequence matters: a Tufts Center for the Study of Drug Development analysis of 106 drugs from 10 firms estimated capitalized pre-approval cost at \$2,558 million per approved drug in 2013 dollars (ideas.repec.org), and BCG states that end-to-end success rates hover around 10% (www.bcg.com). No source reviewed publishes a pharma analogue of Anthropic's index, which is why this report proposes a method and not a benchmark. What exists are adoption surveys, outcome analyses and company-reported use cases, and they differ in sample, definition and incentive.

- **White & Case with Mergermarket (2025).** A survey of 200 life-sciences executives found that 64% use AI regularly in R&D and 88% in human pharma (www.whitecase.com). The human-pharma subsample is only 75 respondents, and only 17% describe their AI strategy as very developed (www.whitecase.com).
- **Pistoia Alliance Lab of the Future (September 2025).** More than 200 experts responded, and 77% expect to use AI within two years, a measure of expected use (www.pistoiaalliance.org).
- **Deloitte 2025 Life Sciences Executive Outlook.** Of 150 C-suite executives surveyed, nearly 60% plan to increase generative AI investment, which is intent measured in 2024 (www.deloitte.com).
- **ISPE Pharma 4.0 survey (2023).** With 418 respondents, AI and machine learning had yet to achieve significant large-scale implementation, though the sample leans toward manufacturing and quality (ispe.org).
- **IQVIA Institute (May 2026).** The Phase I success rate for AI-enabled emerging-biopharma programs was 75%, with Phase II on par with peers, from a cohort the report describes as relatively small (www.iqvia.com).
- **Jayatunga et al. (2024).** AI-discovered molecules showed an 80% to 90% Phase I success rate, and Phase II was comparable to historic averages on a limited sample (pubmed.ncbi.nlm.nih.gov).

Adoption intent is high, stated current use varies sharply with the question, and outcome evidence is early. The 88% and 77% figures are not contradictory, since one measures present use in a small subsample and the other expected use in a broader panel, but they cannot be combined. IQVIA's 75% and Jayatunga's 80% to 90% differ by cohort and method, and both concern discovery-stage AI programs, not agent autonomy. None of these sources rates autonomy level, so none can calibrate an index.

Reported time savings should be labeled by source. MSD reported that an internal generative-AI platform cut the average time to a fully human-reviewed clinical study report first draft from 180 hours to 80 hours (www.msd.com). BCG describes medical writers at one unnamed pharma company who often spent six months producing a single clinical trial protocol (www.bcg.com). Novo Nordisk's case study, published by Anthropic, reports clinical study documentation time falling from 10+ weeks to 10 minutes (claude.com). Anthropic's own announcements place Claude in protocol drafting and, in a later expansion, clinical trial operations and regulatory stages (www.anthropic.com). These are company or vendor reports without independent audit and with different starting definitions. They should not be used to infer an autonomy level, and a draft produced in minutes still passes through human review.

Implications and Future Directions

For governance councils. The most useful import from Anthropic is the structure: map work from records, freeze the map, rate with evidence, weight transparently, report oversight separately, and connect measurements to triggers. The numbers do not transfer. A council that adopts the structure can answer, quarter over quarter, which workflows moved from "assists" to "collaborates" or "leads," and whether oversight kept pace.

For quality and model-risk teams. Link the index to existing risk processes. The credibility file answers the model question, the index answers the delegation question, and the oversight dashboard answers whether monitoring keeps up. One set of tiers across all three prevents contradictions, such as a Tier 3 workflow with a strong credibility file but weak action logging.

For audit. Treat rating evidence, weights and dashboard sources as auditable records.

Open questions. First, rater agreement: with 35% exact human agreement in Anthropic's exercise, pharma rubrics need pilot calibration. Second, third-party verification: no external result is public yet. Third, regulation is moving: the FDA draft remains a draft, EU GMP Annex 22 is a consultation draft, the NIST framework is being revised, and EU AI Act obligations phase in through 2028 (www.fda.gov) (www.nist.gov) (digital-strategy.ec.europa.eu). Tiers should be designed so they can be remapped as final texts appear.

Frequently Asked Questions (FAQs)

How much do AI agents automate pharma research today?

No public study measures this across pharma. Anthropic's 26% "leads" figure applies to its own AI R&D (apnews.com). Pharma surveys measure adoption and intent, with 64% regular R&D use among 200 executives in one survey and 77% expecting use within two years in another (www.whitecase.com) (www.pistoiaalliance.org).

How is AI agent autonomy measured?

It is the degree to which an agent scopes, executes, verifies and acts without human involvement, rated on a scale such as AL0 to AL5; the OECD defines autonomy as the degree to which a system can act without human involvement (oecd.ai), and other work defines levels by the role the user plays (arxiv.org). This report

rates each task from evidence and aggregates by weight.

Is Claude used for drug discovery and clinical trial automation?

Anthropic describes Claude for Life Sciences connectors and use cases including protocol drafting, and a later expansion toward clinical trial operations and regulatory work. These are vendor descriptions. Whether a specific workflow is automated, and to what level, depends on the organization's own evidence.

What human oversight is expected in drug development?

FDA's principles include human-centric design and a clear context of use, and EMA's reflection paper calls for a human-centric approach (www.fda.gov) (www.ema.europa.eu). Under the EU AI Act, oversight for high-risk systems must be commensurate with risks, level of autonomy and context of use (artificialintelligenceact.eu). Requirements depend on the context of use and applicable rules, so confirm with regulatory and quality professionals.

Is there a standard metric for AI agent safety in pharmaceutical research?

No single standard exists. NIST's framework is voluntary and ISO/IEC 42001 specifies an AI management system (www.nist.gov) (www.iso.org). Oversight metrics such as coverage, latency and escalation come from Anthropic's approach and are adapted here.

When should a workflow trigger stronger review or a pause?

When autonomy level, consequence tier and drift together cross predefined limits, as in Table 3. Anthropic suggests such measures could trigger requirements like a fixed testing window.

Conclusion

Anthropic's September 17, 2026 disclosure gave the public a documented method for measuring AI-led work, agent oversight and safety resourcing at a frontier laboratory. It did not provide a pharma benchmark, and the sources reviewed show no independent verification. The method is the contribution: a bottom-up task map, a frozen and versioned basket, evidence-based ratings, transparent weights, and oversight metrics reported on their own.

For pharma R&D, the practical step is to build that scorecard locally. Inventory tasks from SOPs, tickets, notebooks and study plans, protecting confidential and personal data. Rate each task by who scoped, handled exceptions, verified, approved and deployed. Publish the raw level distribution with any weighted index and at least two weightings. Track coverage, latency, escalation, override and rollback separately, with independent sampling to guard against rubber-stamping. Keep model credibility, as framed in the FDA draft guidance, distinct from how autonomously an agent runs a workflow. Finally, tie autonomy, consequence and drift to explicit triggers: enhanced testing windows, independent review, narrower permissions, rollback or pause. Because the regulatory texts are still moving, treat the first quarterly cycle as calibration and re-check conclusions against current official sources.

Source links

Links cited in this article, in order of first appearance. Full addresses are provided for readers using extracted text.

1. <https://www.engadget.com/2261909>
2. <https://apnews.com/article/4d3a7430f57cbc7c39e1c5f2b7d7e132>
3. <https://www.anthropic.com/institute/measuring-pace-of-ai-development>
4. <https://www.fda.gov/media/184830/download>
5. <https://www.whitecase.com/insight-our-thinking/new-frontiers-how-ai-is-transforming-the-life-sciences-industry/state-of-market>
6. <https://www.iqvia.com/blogs/2026/05/iqvia-institutes-global-r-and-d-trends-2026-report-finds-credible-signal-on-ai-enabled-programs>
7. <https://epochai.substack.com/p/toward-an-onet-for-ai-r-and-d>
8. <https://www.theneuron.ai/news/anthropic-claude-leads-26-percent-ai-research-metric>
9. <https://arxiv.org/abs/2506.12469>
10. <https://arxiv.org/html/2506.12469v2>
11. <https://oecd.ai/en/wonk/definition>
12. <https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf>
13. <https://webstore.ansi.org/standards/sae/sae30162021>
14. <https://www.anthropic.com/news/claude-for-life-sciences>
15. <https://claude.com/customers/novo-nordisk>
16. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7400042>
17. <https://pmc.ncbi.nlm.nih.gov/articles/PMC3240751>
18. <https://artificialintelligenceact.eu/article/14>
19. <https://www.anthropic.com/research/measuring-agent-autonomy>
20. <https://arxiv.org/abs/2503.14499>
21. <https://code.claude.com/docs/en/permission-modes>
22. <https://www.anthropic.com/news/claude-science-ai-workbench>
23. <https://www.fda.gov/regulatory-information/search-fda-guidance-documents/considerations-use-artificial-intelligence-support-regulatory-decision-making-drug-and-biological>
24. <https://www.fda.gov/about-fda/center-drug-evaluation-and-research-cder/artificial-intelligence-drug-development>
25. <https://www.fda.gov/about-fda/artificial-intelligence-drug-development/guiding-principles-good-ai-practice-drug-development>
26. https://www.ema.europa.eu/en/documents/scientific-guideline/reflection-paper-use-artificial-intelligence-ai-medicinal-product-lifecycle_en.pdf
27. <https://www.ema.europa.eu/en/about-us/how-we-work/data-regulation-big-data-other-sources/artificial-intelligence>
28. https://health.ec.europa.eu/consultations/stakeholders-consultation-eudralex-volume-4-good-manufacturing-practice-guidelines-chapter-4-annex_en
29. <https://picscheme.org/docview/9715>
30. <https://picscheme.org/en/news/joint-stakeholders-consultation-on-the-revision-of-chapter-4>
31. <https://www.ecfr.gov/current/title-21/chapter-I/subchapter-A/part-11>

32. <https://ispe.org/publications/guidance-documents/gamp-guide-artificial-intelligence>
33. https://database.ich.org/sites/default/files/ICH%20E6%28R3%29_Step4_FinalConsolidatedGuideline_2026_0616_.pdf
34. <https://www.nist.gov/itl/ai-risk-management-framework>
35. <https://www.nist.gov/news-events/news/2026/02/announcing-ai-agent-standards-initiative-interoperable-and-secure>
36. <https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai>
37. <https://ideas.repec.org/a/eee/jhecon/v47y2016icp20-33.html>
38. <https://www.bcg.com/publications/2025/reigniting-biopharmas-research-engine>
39. <https://www.pistoiaalliance.org?p=26823>
40. <https://www.deloitte.com/us/en/insights/industry/health-care/life-sciences-and-health-care-industry-outlooks/2025-life-sciences-executive-outlook.html>
41. <https://ispe.org/pharmaceutical-engineering/september-october-2024/7th-ispe-pharma-40tm-survey-digital>
42. <https://pubmed.ncbi.nlm.nih.gov/38692505>
43. <https://www.msd.com?p=1248010>
44. <https://www.bcg.com/publications/2025/agent-ai-in-biopharma-game-changing-efficiency>
45. <https://www.anthropic.com/news/healthcare-life-sciences>
46. <https://www.iso.org/standard/42001>

THE PUBLISHER

About IntuitionLabs

IntuitionLabs is an AI consulting, custom software development and data engineering firm serving pharmaceutical, biotechnology, medical-device and other life-science organizations. We work with clinical, regulatory, medical-affairs, commercial, quality and IT teams to connect technology decisions with the work people need to accomplish.

AI consulting and adoption

Our AI enablement services cover readiness assessments, use-case selection, governance and policies, team workshops, adoption measurement and ongoing advisory support. We help organizations structure the information layer behind AI: source material, context, permissions and maintained knowledge that make generated answers useful and reviewable. Private LLM inference and hosted AI options support teams evaluating how to operate AI with appropriate control over their data and infrastructure.

Software, data and life-science workflows

IntuitionLabs develops custom software for pharma and biotech, integrates enterprise systems, and builds data engineering and business intelligence solutions. Areas of focus include AI agents, regulatory research, medical writing, medical affairs, CMC information, competitive intelligence and clinical-document workflows. Our eTMF intelligence work includes cross-system reconciliation and inspection-readiness support.

Enterprise platforms and regulated delivery

We provide Veeva services, application support, managed services, integrations and custom applications, alongside enterprise content work involving platforms such as Egnyte. For regulated workflows, our services include GxP enablement, computer-system validation and software development addressing 21 CFR Part 11 requirements. The applicable controls, validation responsibilities and acceptance criteria are defined for each engagement.

Work with IntuitionLabs

Explore [AI enablement](#), [pharma and biotech software development](#), [data engineering and BI](#), and [Veeva services](#). [Contact IntuitionLabs](#) to discuss your workflow, information sources and implementation needs.

IntuitionLabs publishes educational research to help life-science teams make informed technology decisions. Coverage of a product or organization does not imply a client relationship, endorsement or partnership.

Website: <https://intuitionlabs.ai>

This article is educational. Verify consequential medical, legal, financial, regulatory and technical decisions with appropriate professionals and the cited primary sources. Company services are described separately from the article's research findings.