

R&D Evidence Pack

1. The bet meets the record

Engineering to finance, with sub-agents under them, doing the work. Nobody signs off on their commits. The checks do. When they fail, it's public.

Thesis. Our bet: an operating system for every company. People set authority. Checks decide what lands. Software roles, not corporate officers. One GitHub repo. Claude Code seats. Codex builds. Justin hosts workers.

Agents move. Oversight lags.

62% of respondents experiment with or scale agents. No individual function exceeds 10% scaling. McKinsey survey: 1,993 respondents across 105 nations.

Source: [McKinsey, The state of AI in 2025, 2025-11-05](#)

26% of agentic-AI users surveyed cannot detect unauthorized internal agents. EY-commissioned full sample: 202 US senior AI decision makers, public companies, \$1B+ revenue. Subgroup size unstated.

Source: [EY, Autonomous AI implementation outpaces oversight, 2026-09-15 · vendor-reported](#)

Agent seats: configured scope, not current health

Workflow	Mode	Runs	Done
Chief of Staff briefing	operating	40	38
The Machine's own changelog	operating	13	10
Health audit and proposals	operating	13	11
Backlog and roadmap	operating	21	20
The standard and the ratchet	operating	21	20
Legal risk review	operating	20	19
Frontier proposals	operating	8	8
Blog posts, guides and AI Skills entries	operating	16	9
Quarterly and annual reports	operating	2	2
Report and triage	supervised	3	3
Decision queue	supervised	1	1
Advisory panel observers	experimental	10	10

Window: July 21, 2026 to October 22, 2026. Full cadence, authority, boundaries and last completion: [capability matrix](#). Completed is not accepted.

2. Authority has boundaries

Workflow, function, cadence	Status	May do without approval	Boundary
Daily pulse (deterministic, no model) Operations · daily	operating	Measures the organization and files alarms; closes an alarm when its workflow recovers.	Changes nothing it measures.
Rollout ladder evaluator (deterministic) Operations · daily	operating	Promotes or freezes rollout stages from committed evidence.	Rollout decisions follow the evaluator rules, separately from the autonomy supervisor.
Circuit breaker Operations · before each seat run, plus a daily check	operating	Pauses a seat after three consecutive failures; the check runs before the agent starts.	Resuming is a recorded command.
Repair worker (pulse-fix) Engineering · every few minutes on the founder's Mac	experimental	Repairs what the pulse finds. Lands only a short allowlist directly.	Anything on a judge path (gates, protected files, money path) goes to a pull request for the founder.
Handoff loop (multi-session, headless) Engineering · founder-launched	experimental	Works a founder-written queue across sessions, subject to landing checks.	Never touches the founder's calls; stops on a dirty tree, a halt or its weekly budget.
Orbyt Jobs and Orbyt Intelligence code Product development · founder-directed sessions	supervised	AI coding agents (Claude Code, Codex) write the code in sessions the founder directs.	Same gates as every commit; the founder decides what ships.
Autonomy supervisor Executive coordination · daily	experimental	Observes and logs what it would decide.	One defined rule at observe. Promotion requires logged, signed approval; live signed admission is unproven.
Deployment at another company Commercial · phase 2	planned	None yet.	Acceptance: a second company installs and walks a repair with the founder's computer off.
Customer discovery and sales Commercial · not started	human	Customer discovery and sales remain founder-led.	The founder.
Reserved founder decisions Founder authority · as needed	human	Pull-request merges, spending, pricing, legal determinations, protected changes and public claims require founder approval.	Some limits are enforced in code (agent-session valve, signature checks for the unattended workers, pull-request routing); others are instructions to the agents. Only qualifying machine-set scoped halts may be cleared by repair workers. This exception does not apply to the supervisor. Signed admission is not live-proven.

Labels describe configured modes, not current health. operating: runs on its own schedule and lands inside its limits. supervised: runs only when a person dispatches or directs it. experimental: built, not yet proven on live work. planned: not built. human: policy reserves the decision for the founder. Seat workflows can commit permitted output after checks without per-commit approval. Human minutes per workflow are not measured.

3. What is counted

- **Seat run.** One run of an agent seat, recorded by its workflow, not by the agent. 180 rows; 168 counted runs. Excluded from the rate: 8 skips, 4 infrastructure, 0 halted. These are outcome classifications, not proof no work occurred; 12 excluded rows logged turns.
- **Completion.** The output survived the workflow's completion gate and its configured checks. A run the checks refused counts as a failure. **Completion is not acceptance.**
- **Acceptance.** The founder's verdict on an output. 69 disposition rows exist: 2 declared by the founder (2 accepted with edits), 64 derived untouched and 3 derived edited. Untouched means no later edit was found; trusted or unread is not yet measured. 82 completions have no disposition.
- **Founder time.** Reconstructed from his AI session transcripts: prompts clustered by idle time, overlapping windows unioned. The data README specifies the reconstruction method. A floor that misses reading, meetings and work outside those sessions; agent time is never counted as human time. 99% is unclassified. Early months, including July, undercount because of transcript retention.
- **Cost.** Seats use a flat-rate subscription; 172 of 180 rows record tokens. Model tokens are runtime-reported usage, not a dollar cost measure. Subscription, infrastructure and founder costs have not been allocated to individual runs. Dollar cost per run and per accepted outcome has not been measured. These observations do not establish costs for hosted customer use.
- **Unattended versus delegated.** Delegated: the work lands without the founder approving each change, inside limits he set. Unattended: no person started, steered or rescued it. Trigger on record: 64 scheduled, 16 dispatched by a person or process, 100 not recorded (older rows).

How it is computed

Every figure here comes from one snapshot, reproducible from the repository with **node scripts/export-investor-evidence.mjs**. It applies the same classification as the public autonomy ledger and refuses to write if its totals disagree with it. Committed corrections reclassify harness faults first billed to the seat. The sanitized rows, monthly tables, data dictionary and checksums are in the evidence data download.

4. Put the results on the table

Runs by seat

Seat	Runs	Done	Failed	Excl.
chief - of - staff	44	41	3	0
marketing	27	18	9	6
product	21	20	1	1
quality	21	20	1	0
legal	20	19	1	0
engineering	13	11	2	0
innovation	8	8	0	5
operations	5	5	0	0
revenue	5	5	0	0
security	3	3	0	0
finance	1	1	0	0
Total	168	151	17	12

By seat, not by workflow: a seat can run more than one workflow (Chief of Staff also runs the weekly changelog), so these rows can exceed page 1's. Check-pass rate 89.9%. Early subset of this ledger: 80.0% (30 runs, frozen 2026-08-11). Logging is incomplete; at least one known run is missing. No independent validation.

By month

Month	Done	Failed	Rate	Excl.
2026-07	13	2	86.7%	0
2026-08	40	8	83.3%	0
2026-09	52	5	91.2%	5
2026-10	46	2	95.8%	7

Failures

17 counted failures: 11 runs that failed, 6 refused by the checks, 3 with caller/isolate and guard-refused recorded, not the individual check. Partial months: July 21 to 31; October 1 to 22. Skips ("nothing to do") and harness faults sit outside. Count faults as failures: October 46 of 49, 94%. Gates changed: September 20, no_sentinel retired; September 23, 3 faults reclassified. Checks changed; rates do not measure constant-quality work.

5. Agents bluff. Checks don't.

September 22, 2026. Agent claimed done. Check refused.

04:37 · State changes	Operations committed new state.
Claimed done	Engineering called it complete.
Refused	Stale file on main, not seat code. Still counts against the seat.
05:43 · Fix it	Founder-launched loop: replay, fix, regression test. No intervening founder prompt. CI passed.
06:43 · Pass the checks	Engineering re-dispatched. Checks passed. Report landed. Dispatcher unrecorded.
Human hand	Justin launched it and wrote its queue. A separate defect later stopped it.

The refusal got a record

A change landed September 14 that records a check's refusal beside the agent's claim of completion. Before it, refused runs carried no reason. It has since named the cause of 3 production refusals; this morning's is one.

The proof stops here

- Shows: a founder-launched loop repaired a failure from its founder-written queue, with no intervening founder prompt. Engineering was dispatched again and passed checks. The loop later stopped on a separate defect.
- Does not show: how often this works, how much founder time the system needs over months, or that an independent reviewer would accept the repair. One case shows the mechanism worked once.

Full case record with UTC times: <https://www.orbytlabs.ai/investors/data/case-2026-09-22.md>. Run ids, commit hashes, the CI run and the transcript lines are available on request.

Founder active time (floor)

Month	Hours	Classified
2026-07	1.4	0.0
2026-08	76.9	1.3
2026-09	175.4	3.1
2026-10	116.5	0.0

At least 370 hours over July 6 to October 19, 2026. Source: founder-effort-monthly.csv. Classified hours are assigned by rule; review has no rater. Early months undercount through transcript retention.

6. The refusal stays. The repair follows.

We kept the refusal on record

The September 22 record concerns our own workflow machinery. A stale public file on main made the engineering check fail. The seat's code did not cause it. The refusal remains a counted failure because the record has no category that excuses it.

- **What changed.** The founder-launched loop repaired the workflow and added a regression test. Engineering was dispatched again and its report landed after checks.
- **What stayed unproven.** Repeatable unattended recovery, independent acceptance and reduced founder supervision.

The repair cleared its checks

In early October repeated walks of the repair worker exposed missing sandbox logs, a turn limit inside its own commit hook and a test that killed its own process tree. Each defect was traced and fixed. A later attempt landed: its change touched a protected path, so the landing gate sent it to a pull request; the founder reviewed and merged it, and the next pulse reproduced all recorded run configurations after previously reporting a contradiction. A planted regression with a false claim of a fix was then shown still red. Founder time is recorded on the milestone. The record labels it a first version, pending a stricter before-and-after check.

Put the failures to work

97 failure modes are documented in public, 73 with a countermeasure in code. By kind: 53 verification gap, 23 operator process, 21 external behavior. Corpus: mostly founder-directed sessions. About one in four touch our machinery. No origin field. 11 undated; 6 predate the run window. 106 of 168 counted runs: CoS, quality, product and legal reports. Marketing: 18 of 27 passed; 9 of 17 failures. Minutes saved: unmeasured.

Model tokens (millions)

Month	Runs	Output	Cache read
2026-07	7	0.20	26.28
2026-08	48	1.58	265.53
2026-09	62	2.36	525.29
2026-10	55	1.88	284.44

Source: seat-tokens-monthly.csv. Runtime-reported usage only. Subscription, infrastructure and founder costs are unallocated; hosted economics are unproven.

7. Make the bet face the tests

Competition and platform risk

GitHub Agent HQ Multi-agent control, permissions and audit for repository work.

Source: [GitHub, 2025-10-28 · vendor-reported](#)

OpenAI AgentKit October 2025 launch. June 2026 update: Agent Builder and Evals end November 30, 2026; other OpenAI tools recommended.

Source: [OpenAI, 2026-06-03 · vendor-reported](#)

CrewAI Agent orchestration and an enterprise platform in its October 2024 announcement.

Source: [CrewAI via GlobeNewswire, 2024-10-22 · vendor-reported](#)

Microsoft Agent 365 Agent registry, access control and quarantine for enterprise governance.

Source: [Microsoft, 2025-11-18 · vendor-reported](#)

LangChain / LangGraph LangChain and LangGraph: agent frameworks and orchestration. LangSmith: tracing and evaluations.

Source: [LangChain, 2025-10-20 · vendor-reported](#)

Claude Agent SDK An SDK using Claude Code's agent loop and tools.

Source: [Anthropic \(guide reviewed\), 2026-10-20 · vendor-reported](#)

Where the bet starts

73,499 sums two Census 2022 counts below 100 employees: 12,610 software publishers plus 60,889 custom programming firms. Possible double counting; not unique prospects or measured demand.

Source: [US Census Bureau, SUB 2022, 2025-04-10](#)

The commercial bet

Plan: Apache-2.0 open core. Paid hosting. Release scope and branding remain under review. **Proposed:** Open: seats, workflows, checks. Paid: cross-repository history, governance evidence, support. **Pricing:** Unset. Informed by token usage. Plans as of October 22, 2026, not promises. Installation, demand, supervision savings and economics unproven. Failure risks: technology, founder/provider dependence, costs, competition, unresolved branding rights.

8. The proof still owed

Bring a harder test.

Codex drafts from written instructions; Claude reviews. Bring harder research questions. No provider endorsement or independent validation.

Each limit and the test that closes it

- Acceptance: earn it. 64 of 67 measured outputs untouched by the founder; 3 derived edited. Trusted or unread is not yet measured. 82 completions have no disposition. Test: record reads and verdicts.
- Founder time: at least 370 reconstructed hours; 99% unclassified. Early months undercount: transcript retention. Test: classify two weeks.
- Economics: Model tokens are runtime-reported usage, not a dollar cost measure. Subscription, infrastructure and founder costs have not been allocated to individual runs. Dollar cost per run and per accepted outcome has not been measured. These observations do not establish costs for hosted customer use. Test: allocate costs per accepted output.
- Outside: comparison, external installation, demand unproven. Hosting unfinished; not for sale. Tests: matched work, second-operator repair, buyer interviews.
- Provider dependence: Scheduled agent roles currently depend on Claude Code. Provider access, usage limits, terms and pricing may change. The recorded setup does not establish permission or economics for a hosted customer service. Providers can compete. Test: permissions review and another runtime.
- Logging gaps. Detailed records: 107 of 180 rows. Test: reconcile dispatches.
- Authority: code enforces paths, landing checks, HALTs. Spending, pricing, legal calls and claims rely on instructions and judgment. Signed admission is coded, not live-proven. Test: protected actions and signed restarts.

Next tests, proposed. Does it pay for itself? Count results: Acceptance. Supervision minutes. Allocated costs. Make it compete: Same work, models, tools. Coordination versus none. Hand it over: Another operator installs and repairs. Founder's computer off. Face the buyer: Willingness to pay. Recurring work. 6 briefs. Single agent: 1 resolved. 3-agent team: 0. Extra roles retired. Agent-reviewed; human time unmeasured. Role splitting only, not the whole system. Learning-study counts are not yet verified.

Justin Bartak, Founder and CEO

hello@orbytlabs.ai · <https://www.orbytlabs.ai/investors>

[Research or partnership conversation](#)