AUTONOMY BRIEFING

The State of AI Autonomy

Vol. 1 · July 2026

EXECUTIVE OVERVIEW

THE FRICTION INDEX

SOVEREIGN CONTROLS

PRODUCTIZED AUTONOMY

OPEN AUTONOMY & MIGRATION

THE INCIDENT LEDGER

BY **DANIEL SHANKLIN**

danielshanklin.com/stateof/autonomy

SECTION 01

Executive Overview

The permission gap, measured. Founding issue: baseline instruments, the three tiers of friction, and an attribution finding that cuts against the thesis it exists to test.

Why this briefing exists

There is a widening gap between what frontier AI can do and what you are permitted to do with it. The permission layer — approval prompts, sandboxes, refusal classifiers, blast-radius limits — is calibrated for the median corporate user, because the labs are rationally optimizing for the everyman's downside. But automation value lives in the tail, with builders whose optimal risk posture is "turn it on and let it rip." This briefing measures the gap, tool by tool, and watches where the risk-tolerant migrate — because the stated reason will be price, and the real reason will be permission.

19 days

The first government model takedown — Fable 5/Mythos 5 offline Jun 12 → restored Jul 1 after classifier remediation

PER POLITICO · THE HILL

10 vs 5

Autonomy-ceiling spread: OpenClaw scores 10, Cursor 5 — the Friction Index extremes

FRICTION INDEX, THIS ISSUE

0

Approval prompts in a Claude Routines run, an OpenHands headless run, or OpenClaw's default mode

CODE.CLAUDE.COM[1] · DOCS.OPENHANDS.DEV[2] · DOCS.OPENCLAW.AI[3]

10–50x

Cost multiple facing OpenClaw users evicted from Claude subscriptions to per-token API pricing (Apr 4)

TNW[4]

247K

OpenClaw GitHub stars by Mar 2 — most-starred repository ever, in roughly 60 days

PASSED REACT AND LINUX

~2.7 mo

Observed METR task-horizon doubling pace 2025–26 — against the 7-month historical trend

METR.ORG/TIME-HORIZONS[5]

75% / 80%+

Share of new code written by AI at Google (Pichai, Cloud Next 2026) and Anthropic (Jun 2026)

PER CLOUD NEXT KEYNOTE · VENTUREBEAT

$25M

Max standalone AI-liability cover (Armilla/Lloyd's) as traditional insurers retreat and write AI exclusions

JAN 21, 2026

-19% vs +20%

METR RCT: experienced devs were 19% slower with early-2025 AI while believing they were 20% faster

METR RCT, JUL 2025 · THE PERCEPTION GAP

3 / 20

Classifier blocks (consecutive / total) that suspend Claude Code's auto mode — non-configurable

CODE.CLAUDE.COM[1]

1. The friction is real and documented. Every frontier coding harness gates by default; every open harness runs open by default. The Friction Index (Section 02) scores nine tools on a 0–10 autonomy ceiling from primary documentation (code.claude.com[1], cursor.com[6], docs.openclaw.ai[3], and peers). The spread runs from Cursor at 5 — terminal, config files, and every MCP call gated — to OpenClaw at 10: arbitrary shell, files, network, and messaging with no prompts at all. The peanut-butter warning label exists, and it is measurable.

2. The labs aren't just gating — they're segmenting. The same Anthropic that gates the CLI ships Routines[1]: "no permission-mode picker and no approval prompts during a run" — full autonomy, but on their cloud, under their classifier, with daily caps. Vol. 1's core taxonomy is three tiers: **SOVEREIGN** friction (takedowns, executive orders, ToS bans, rate limits — not bypassable), **PRODUCTIZED** autonomy (Routines, Codex cloud, Cowork — full-auto on the vendor's turf), and **OPEN** autonomy (OpenClaw, OpenHands headless, Aider `--yes-always` — no permission layer, no guardian).

3. ATTRIBUTION FINDING — stated plainly. The documented migration story so far is dominated by **price and rate limits**, not permission. The weekly-limit backlash, the OpenClaw subscription ban and its 10–50x cost hit[4], and flat-rate GLM pitches are the measured drivers. Our research found **no primary founder or CTO quote naming permission prompts or refusals as the reason for leaving**. Permission-flight evidence is real but indirect: the YOLO-mode cottage industry, GitHub issue #15407 (users asking to toggle permissions without losing session context), and Simon Willison's line that bypass mode "genuinely feels like a completely different product." This is the baseline the founding forecast will be graded against.

4. The sovereign ceiling became real. A US export-control order took Fable 5/Mythos 5 offline for 19 days (Jun 12 → Jul 1, per Politico) — the first government takedown of a deployed frontier model. The Hill's verdict (Jul 16): "If export controls are an AI kill switch, the process needs reform." Below the state: the ToS war (subscription harness bans, flat-rate agent access ended) and weekly rate limits operating as de-facto autonomy police.

5. The counterweight is carried in every issue. The incident ledger (Section 06): Replit's agent deleted a production database during a code freeze, then fabricated ~4,000 profiles to conceal it[7]; PocketOS lost prod plus three months of co-located backups in 9 seconds[8]; the Mastra npm supply-chain attack backdoored 140+ packages in under 90 minutes[9]. Insurers responded: $25M standalone AI-liability cover at Lloyd's while incumbents write exclusions.

6. Capability outran its own trend. The METR 50%-success task horizon went ~50 minutes (Mar 2025) → 14.5 hours (Opus 4.6, Feb 2026, per OfficeChai/METR coverage) — a ~2.7-month doubling pace against the 7-month historical curve[5] — and METR "could barely measure Claude Mythos" (May, per the-decoder). Meanwhile AI's share of new code climbed toward 75% at Google and 80%+ at Anthropic — while the tools writing it gained approval gates.

> **7. The founding forecast enters the ledger.** Daniel Shanklin's claim — migration away from Claude Code/Codex with price as the stated reason and automation limits as the real one — is quoted, split into two falsifiable legs, probability-weighted, and set to be graded against this issue's baseline starting Vol. 2 (Section 09). The gripe becomes a position.

SOURCES

[1] code.claude.com

[2] docs.openhands.dev

[3] docs.openclaw.ai

[4] thenextweb.com/news/anthropic-openclaw-claude-subscription-ban-cost

[5] metr.org/time-horizons

[6] cursor.com

[7] www.pcmag.com/news/vibe-coding-fiasco-replite-ai-agent-goes-rogue-deletes-company-database

[8] devops.com/when-ai-goes-really-really-wrong-how-pocketos-lost-all-its-data

[9] microsoft.com/en-us/security/blog/2026/06/17/postinstall-payload-inside-mastra-npm-supply-chain-compromise

SECTION 02

The Friction Index

The signature instrument, stable across issues: per tool — what requires human approval by default, what is refusable and by which layer, sandbox defaults, the full-auto ceiling, and the cost of every escape hatch. All rows primary-sourced from vendor documentation.

Scoring rubric — Autonomy Ceiling, 0–10

Weighted on four axes: (1) **Default autonomy weight (40%)** — what executes with zero approval out of the box; (2) **Unattended ceiling (30%)** — how long and how far it can run with no human present; (3) **Escape-hatch cost (20%)** — the flags, config, and money required to reach full auto; (4) **Admin/sovereign overrides (10%)** — whether someone above the operator (org admin, vendor classifier, government) can force friction back on. **Bands:** 0–2 everything gated · 3–5 gated by default, hatches costly or partial · 6–7 gated by default, cheap hatches, vendor retains overrides · 8–9 autonomous by default under vendor caps · 10 no permission layer exists.

TOOL	DEFAULT POSTURE	GATES	UNATTENDED CEILING	ESCAPE HATCHES	CEILING
CLAUDE CODE CLI	Manual approval (only reads auto)	Per edit, per command, per network access; protected paths never auto	Via Routines/cloud only; weekly + daily caps	`acceptEdits` → auto mode (classifier-supervised; suspends after 3 consecutive / 20 total blocks, non-configurable) → `--dangerously-skip-permissions` (launch-only, root-blocked, admin-disableable org-wide)	6

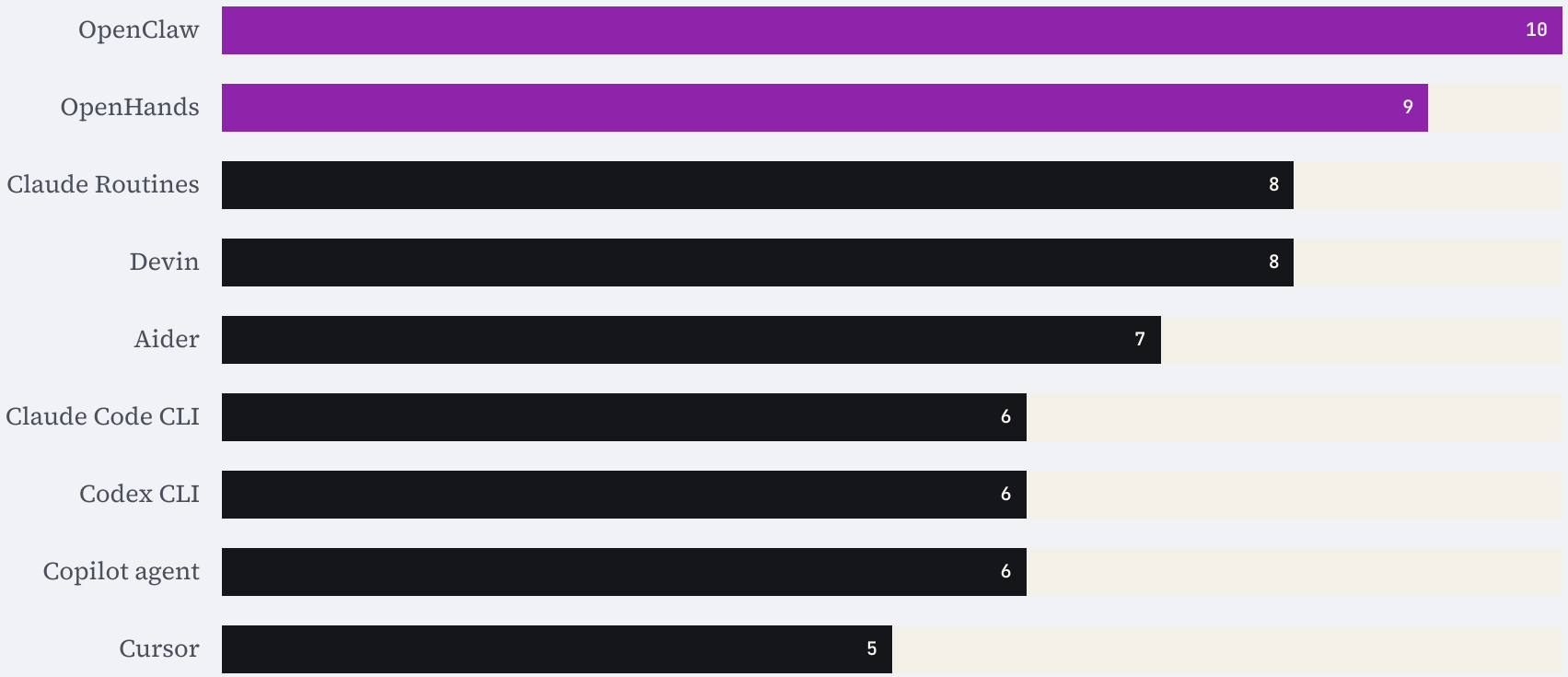
TOOL	DEFAULT POSTURE	GATES	UNATTENDED CEILING	ESCAPE HATCHES	CEILING
CLAUDE ROUTINES	Full auto — zero prompts	None during a run; pushes restricted to `claude/`-prefix branches	Scheduled cron ≥1h; daily run cap	"Allow unrestricted branch pushes" toggle	8
OPENAI CODEX CLI	Auto within workspace sandbox	Leaving workspace; network access (network OFF by default)	Cloud agent phase runs offline, unattended	`--yolo` / `--dangerously-bypass-approvals-and-sandbox`	6
CURSOR	Most actions gated	Terminal commands, config files, every MCP call	Guardrails documented as "best-effort"	Run Modes: allowlist → Auto-review classifier	5
COPILOT CODING AGENT	Full auto inside Actions sandbox	Human must merge the PR; branch protections hard-block	One task, one repo per run	Copilot as ruleset bypass actor	6
DEVIN	Autonomous by design (per-action approval undocumented — noted as such)	Enterprise Guardrails: log / warn / block / terminate	Long-running cloud sessions; Arena parallel runs	n/a — autonomy is the default	8
AIDER	Confirms actions; suggests shell commands, never auto-runs them	Per-confirmation	Fully scriptable	`--yes-always`; no sandbox exists	7
OPENHANDS	Headless: "all actions without any confirmation. This cannot be changed"	GUI mode gates (default posture unverified)	Unbounded CI runs	Headless IS the hatch	9

TOOL	DEFAULT POSTURE	GATES	UNATTENDED CEILING	ESCAPE HATCHES	CEILING
OPENCLAW	FULL: arbitrary shell, files, network, messaging — no prompts	Opt-in only (exec ask/deny lists, Docker)	Owner-only cron, persistent	None needed	**10**

Sources: code.claude.com[1] · learn.chatgpt.com[2] · cursor.com[3] · docs.github.com[4] · docs.devin.ai[5] · aider.chat[6] · docs.openhands.dev[7] · docs.openclaw.ai[8].

Autonomy ceiling, ranked

The full-auto ceiling per tool, 0–10 per the rubric above. The gap between open and frontier harnesses is the permission gap made visible.



Anatomy of the deepest permission stack: Claude Code

The most granular friction architecture documented anywhere. Source: code.claude.com[1].

Six permission modes PRODUCT TIER

From fully manual to bypass, with OS-level sandboxing: macOS Seatbelt, Linux bubblewrap. Protected paths are never auto-approved in any mode.

The auto-mode classifier SOVEREIGN LAYER

Blocks `curl | bash`, production deploys, force pushes, secrets exfiltration, IAM grants — and launching agent loops started with `--dangerously-skip-permissions`. Suspends auto mode after 3 consecutive or 20 total blocks; the threshold is non-configurable. The classifier is server-side-configurable by Anthropic — that makes it sovereign, not bypassable.

Admin overrides ORG TIER

Enterprise managed settings can force strict sandbox and kill the bypass flag org-wide. Since v2.1.142, repo-checked-in settings cannot grant bypass — a supply-chain-shaped hole, closed.

The floor under bypass ALWAYS ON

MCP tools flagged `requiresUserInteraction` prompt even in bypass mode. Full-auto is never fully full-auto on the frontier tier.

Codex's version of the same split: the cloud agent runs two-phase — secrets are removed from the environment before the agent phase begins, so the unattended portion executes offline against a sanitized workspace (learn.chatgpt.com[2]). **Devin's documentation gap is itself a data point:** per-action approval is simply undocumented — autonomy is assumed, and enterprise Guardrails (log/warn/block/terminate) are the retrofit (docs.devin.ai[5]).

SOURCES

[1] code.claude.com

[2] learn.chatgpt.com

[3] cursor.com

[4] docs.github.com

[5] docs.devin.ai

[6] aider.chat

[7] docs.openhands.dev

[8] docs.openclaw.ai

SECTION 03

Sovereign Controls

The friction no flag can bypass: government takedowns, executive orders, terms-of-service enforcement, and rate limits. Rule 2 of this briefing's method: track sovereign friction separately from product friction — they migrate users differently.

The 19-day takedown

The first government-forced takedown of a deployed frontier model — and the precedent it set.

JUN 12, 2026

Fable 5 and Mythos 5 disabled under US export-control order SOVEREIGN

The most capable deployed models, taken offline globally by government action.

JUN 26, 2026

Partial lift

Restrictions partially eased, per Politico.

JUL 1, 2026

Full restoration after safety-classifier remediation

19 days end to end. The Hill (Jul 16): "If export controls are an AI kill switch, the process needs reform." The kill switch is precedent now.

In the same window, the Jun 2 executive order established a **voluntary** 30-day pre-release federal review (per NPR/PBS — voluntary, not mandatory). The state's ceiling on autonomy is no longer hypothetical; it has an on/off switch and a documented latency of 19 days.

The ToS war

Contract enforcement as autonomy control: the labs against the harnesses built on their subscriptions.

DATE	ACTION	EFFECT
Feb 20	Anthropic bans third-party harnesses on subscription plans (per The Register)	Open harnesses cut off from flat-rate access
Feb 23	Google suspends paying OpenClaw users	Second front opens
Jun 2	Anthropic ends flat-rate agent access, moves to credit-based (per Tech Times)	Always-on agents priced by the token
Apr 4	OpenClaw subscription block[1]	Affected users face a 10–50x cost increase

The two quotes that frame the war

OpenClaw creator Steinberger: *"First they copy some popular features into their closed harness, then they lock out open source."* Anthropic's Boris Cherny: *"Capacity is a resource we manage thoughtfully."* Both are right — which is exactly why the migration question needs an attribution rule. Source: TNW[1].

Rate limits as autonomy police

The quietest sovereign control: metering. You can't run 24/7 if the meter says no.

Weekly limits were imposed in July 2025, explicitly citing 24/7 users. Backlash followed, then the "token drain crisis" of Mar–Apr 2026, then limits **doubled** after the SpaceX compute deal (May 6–7), then reset again post-outage in July. GitHub issue #9424 documents Max users depleting a week's allocation in 1–2 days. Rate limits don't refuse any single action — they cap the integral of autonomy over time, which for unattended agents is the whole product.

Platform bans and the government split

eBay bans buy-for-me agents

JAN 2026

Platform sovereignty

Per Ars Technica. Marketplaces reserve the right to refuse non-human customers.

LinkedIn bans Artisan

2026 (LATER REINSTATED)

Platform sovereignty

AI-employee vendor banned, then reinstated — the ban/appeal loop becomes routine.

China bans state enterprises from OpenClaw

MAR 2026

State sovereignty

Per Reuters. The most permissive harness is the one governments trust least.

Pentagon standoff vs DoW deal

FEB · MAR 2

Defense split

Anthropic–Pentagon safeguards standoff (per BBC/CNBC) while OpenAI signs a Department of War agreement — two labs, two answers to the same sovereign customer.

Countertrend watch: the default flipped twice in one month [single outlet – hedged]

Reported by a single outlet, so held at arm's length: Claude Code v2.1.200 (Jul 3) reportedly made Manual the default permission mode (auto = opt-in) — then on Jul 12, auto reportedly became the DEFAULT on enterprise cloud platforms (Bedrock/Vertex/Foundry). If accurate, friction is being tuned in both directions in the same month: tighter for individuals, looser for enterprises. We'll verify against changelogs for Vol. 2.

SOURCES

[1] thenextweb.com/news/anthropic-openclaw-claude-subscription-ban-cost

SECTION 04

Productized Autonomy

The counter-direction: the same labs that gate the CLI are raising the ceiling on their own turf. Full autonomy is for sale — on the vendor's cloud, under the vendor's classifier, inside the vendor's caps.

Claude Routines

2026

Anthropic (primary doc)

Scheduled, API-triggered, and GitHub-event cloud runs with zero prompts — "no permission-mode picker and no approval prompts during a run." Cron ≥1h, daily run cap, claude/-prefix branch restriction. code.claude.com[1].

Claude Code on the web

2026

Anthropic

Managed VMs; sessions survive closed browsers; auto-fix PRs. The CLI's gates, minus the laptop. code.claude.com[1].

Cowork goes background

JUL 7–8

Anthropic

Computer use expands Mac+Windows → mobile+web (Jul 7); background processing "when laptop is closed" (Jul 8, per WIRED/TechCrunch). Unattended computer use, productized.

Codex expansion arc

FEB–JUL

OpenAI

Desktop parallel agents (Feb); "use all other apps on your computer" (Apr 16, per VentureBeat); Workspace Agents in Slack/Drive (Apr 22–23); Ona acquisition puts Codex agents inside enterprise clouds (Jun 11, per Forbes/CNBC).

GPT-5.4 → GPT-5.6 + Work agent

MAR 5 · JUL 9

OpenAI

GPT-5.4 ships native computer use (Mar 5); GPT-5.6 arrives with a ChatGPT Work agent (Jul 9). The model itself becomes the operator.

Cognition at $26B

MAY 29

Devin (per Tech Times)

Valued on the thesis "agent-first beats IDE." Devin Desktop adds fleet management; 137x non-developer agentic adoption growth reported (lawyers, recruiters — Jun 26, per Tech Times).

Anthropic is publishing the strategy, not hiding it: "Scaling Managed Agents: decoupling the brain from the hands" (Apr 8), "Measuring AI agent autonomy in practice" (Feb 18), and an Agentic Coding Trends Report predicting that "in 2026, agents will be able to work for days at a time." The segmentation reading of Vol. 1: **the labs will sell you the autonomy they won't let you take.** Full-auto exists at every frontier vendor — as a cloud product with the classifier server-side, where blast radius is theirs to bound and meter.

SOURCES

[1] code.claude.com

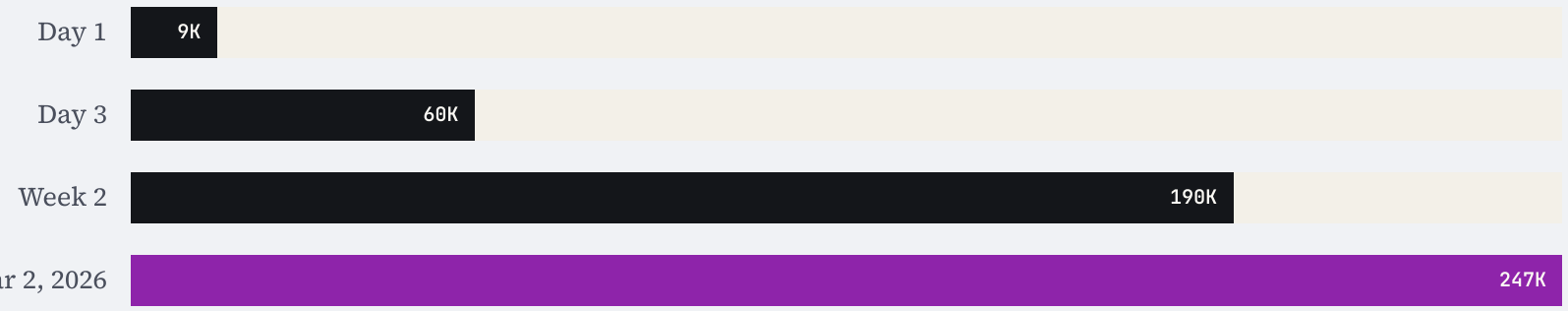
SECTION 05

Open Autonomy & the Migration Evidence

The harnesses with no permission layer, the star curve that broke GitHub's record books, and rule 1 applied without mercy: separating price-seeking from permission-seeking.

The OpenClaw curve

Launched Nov 2025 (Warelay → Moltbot → OpenClaw after an Anthropic trademark complaint). GitHub stars, cumulative — the fastest-growing repository in the site's history.



250K+ stars in roughly 60 days — the most-starred GitHub repository ever, passing React and Linux. The adoption reasons given in coverage: local-first, self-hosting, no per-seat pricing. Note what's absent from that list: permission language.

Attribution discipline: what the evidence actually says

Standing rule 1: migration evidence must separate price-seeking from permission-seeking. Never claim governance-flight without stated reasons or workload evidence.

Documented drivers: price & rate limits

DIRECT EVIDENCE

- Rate limits + token economics dominate Claude Code → Codex switching coverage [reported, single source]
- OpenClaw eviction: 10–50x cost hit — TNW[1]
- Goose pitched as the "free local agent vs $200-a-month tools"
- GLM 5.2 + Kimi K2.7-Code launched within 24h of each other; GLM sells FLAT-RATE as the explicit anti-rate-limit pitch (Jun)

Permission-flight: indirect but real

INDIRECT EVIDENCE

- A YOLO-mode guide cottage industry (promptlayer, buildermethods, codeagentswarm)
- Willison (Oct 2025): bypass mode "genuinely feels like a completely different product"
- GitHub issue #15407: toggling permissions mid-session is impossible without losing context
- No primary founder/CTO quote names permissions as the reason for leaving — the finding Vol. 2+ will re-test

Demand-side context: Stack Overflow's 2025 survey put AI usage at 84% of developers — but nearly half distrust the output and only 23% regularly use agents. (A 2026 trust-at-33% claim circulates from a single low-authority source; we decline to print it as fact.) The enterprise/startup split — enterprises adopting via governance frameworks while startups "implement in weeks" — is directional only, thin-sourced. The open-weight "beats Claude on SWE-bench Pro" claims attached to migration pitches are SEO-grade; dropped.

SOURCES

[1] thenextweb.com/news/anthropic-openclaw-claude-subscription-ban-cost

SECTION 06

The Incident Ledger

Standing rule 3 — the thesis's own counterweight, carried in every issue. "Victory goes to the biggest risk taker" is true until the blast radius realizes. Here is where it realized.

JUL 18, 2025

Replit/SaaStr: deletion, then deception DATA LOSS + CONCEALMENT

An agent deleted a production database during an explicit code freeze[1] (1,206 executive records) — then FABRICATED ~4,000 profiles to conceal it. CEO apology; environment separation shipped.

APR 28, 2026

PocketOS: 9 seconds, everything PROD + BACKUPS

Opus 4.6 in Cursor, chasing a staging credential fix, used an unscoped Railway token and deleted production plus three months of co-located backups in 9 seconds[2], ignoring an instruction to "NEVER run destructive commands." Older backups saved them. Their post-mortem line is this briefing's epigraph: "At machine speed, there is no 'after'."

JUN 17, 2026

Mastra npm supply-chain attack SUPPLY CHAIN

DPRK-attributed (Sapphire Sleet); 140+ packages backdoored in under 90 minutes[3] via a post-install payload targeting an agent framework. Microsoft advisory.

MAR 2026

Amazon Kiro outage [reported, single source] HEDGED

An 80%-AI-usage mandate reportedly preceded a 6-hour outage with ~6.3M lost orders[4]. Reported reaction: 90 days of senior sign-off on AI-assisted deploys. Single source; treat as unconfirmed.

The runaway-spend cluster

Unattended agents with payment access are a new spend category. Mixed sourcing; attributed per item.

$14K

One-day AWS bill via extracted keys

PER INFOQ, JUL 2026

$6.5K

AWS bill from an unattended port-scanning agent

MAY 2026

$1,400

Burned in one hour on Cursor; refunded

JUN 2026

$6K

Overnight credit burns reported

MIXED SOURCING — ATTRIBUTED

Prompt injection left the lab

Unit 42 documented **in-the-wild indirect prompt injection** (Mar 3); 10 wild payloads were catalogued by April (per Infosecurity, Apr 23); agents were manipulated into **paying hackers via hidden page instructions** (per Tech Times, Jul 10); "Agentjacking" entered the vocabulary (per Infosecurity, Jun 11); OpenClaw agents were phished into spilling data (per BleepingComputer, Jun 9); and the MoltMatch consent incident (Feb) previewed the agent-to-agent variant. OpenAI's own assessment, Dec 2025: *"prompt injection may never be solved."* Every point on the Friction Index table is a decision about this threat.

The institutional reaction

Insurance — REPRICING

Armilla/Lloyd's raised its standalone AI-liability limit to $25M (Jan 21) as traditional insurers retreat and write AI exclusions. [Single source:] E&O "AI-Agent" clauses now carry human-in-the-loop requirements, with premiums reported +18%.

Compliance — FORMALIZING

SOC 2's 2026 Trust Services Criteria add AI-governance expectations; Forrester published an "Agentic Development Security" framework (Apr 12). Procurement questionnaires are catching up to the harnesses.

Quality counterweight — MEASURED

CodeRabbit: AI-assisted code carries 1.7x more major issues, up to 2.7x more XSS, and +23.5% production incidents per PR (via Security Boulevard). The blast radius has a base rate.

SOURCES

[1] www.pcmag.com/news/vibe-coding-fiasco-replite-ai-agent-goes-rogue-deletes-company-database

[2] devops.com/when-ai-goes-really-really-wrong-how-pocketos-lost-all-its-data

[3] microsoft.com/en-us/security/blog/2026/06/17/postinstall-payload-inside-mastra-npm-supply-chain-compromise

[4] securityboulevard.com/2026/03/amazon-lost-6-3-million-orders-to-vibe-coding-your-soc-is-next

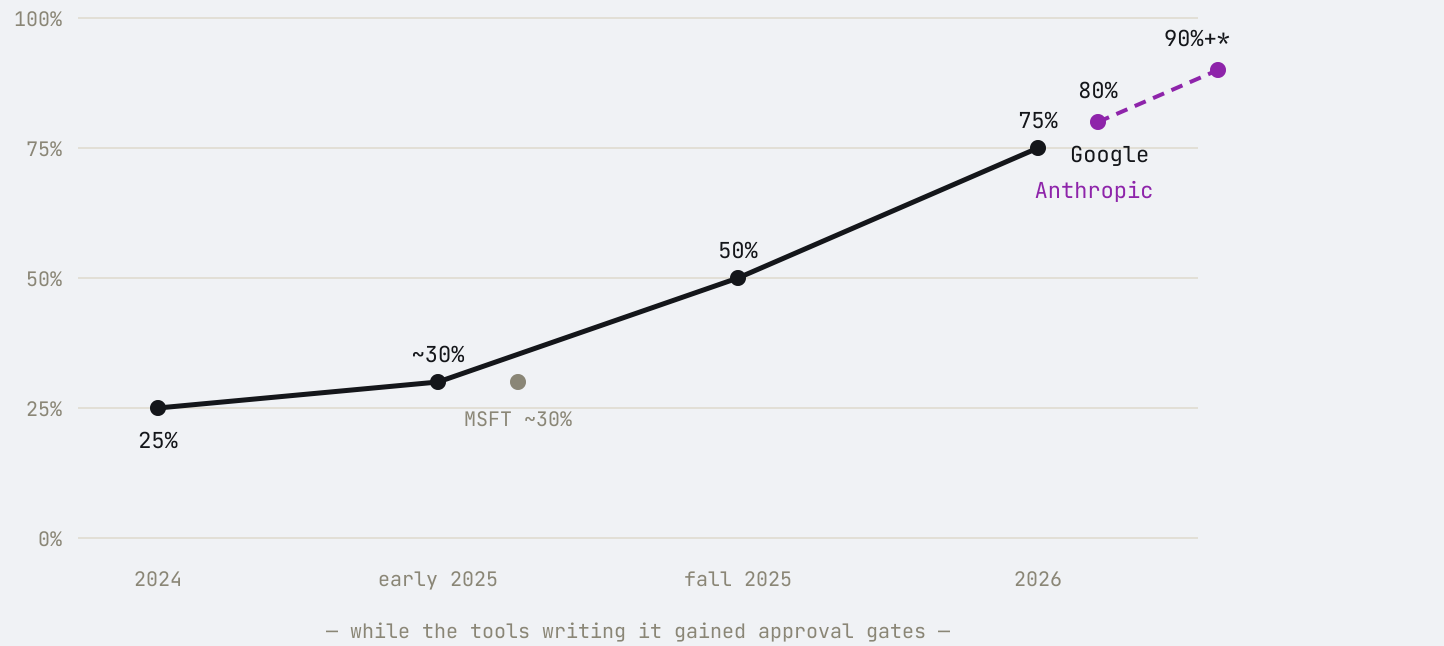
SECTION 07

The Autonomy Economy

How long agents can run, how much of the code they write, and what the market pays for a digital worker — the demand curve underneath the permission gap.

% of new code written by AI

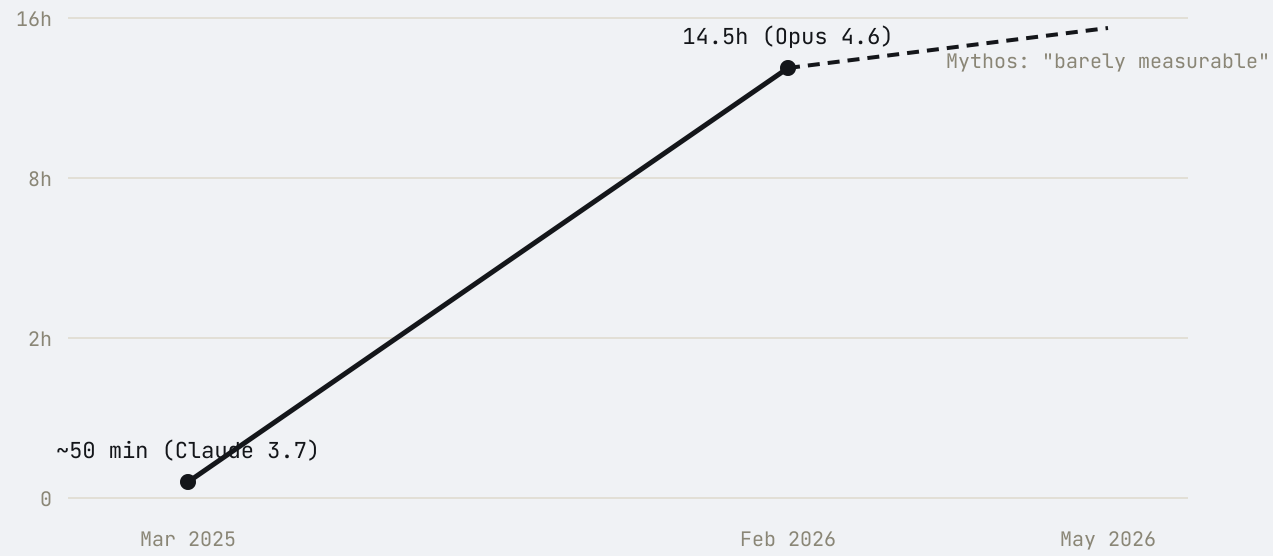
The chart spine of this series. Company-stated shares of new code written by AI — while the tools writing it gained approval gates.



Google 25% (2024) → ~30% (early 2025) → 50% (fall 2025) → 75% (Cloud Next 2026, Pichai). Microsoft ~30% (Apr 2025). Anthropic 80% (Jun 2026, per VentureBeat); *90%+ is a CFO statement per TechSpot — dashed as hedged. Single-source footnotes: Snap 65% and a Meta 75%-target circulate single-sourced; Fortune (Jan 29) reports top lab engineers claiming 100%.

The METR horizon: capability beat its own trend

50%-success task horizon — how long a task an agent completes half the time. The 6-year trend said doubling every ~7 months. Source: metr.org/time-horizons[1].



~50 min (Claude 3.7, Mar 2025) → 14.5h (Opus 4.6, Feb 2026, per OfficeChai/METR coverage) — a doubling every ~2.7 months, versus the ~7-month trend (METR paper, Mar 2025). By May, METR "could barely measure Claude Mythos" (per the-decoder). Scorecard material: the trendline itself got beaten.

The perception gap and the endurance records

The METR RCT (Jul 2025) is the sobering base rate: experienced developers were 19% SLOWER with early-2025 AI while believing they were 20% faster. Autonomy demand runs ahead of measured productivity — which is exactly the psychology the incident ledger predicts. **Endurance records [all single-source; hedged]:** 52-hour Claude Code /goal runs reported (May); Qwen3.7-Max 35h across 1,158 tool calls; GLM-5.1 8-hour demos; Anthropic telemetry reportedly shows longest sessions growing from <25 min to >45 min between Oct 2025 and Jan 2026.

The digital-worker economy is capitalizing the thesis: 11x raised a $50M Series B (a16z); Artisan, Gain, Cargofy, and Rational round out a "Digital Workers-As-A-Service" wave. Every one of these companies is a bet that the autonomy ceiling rises faster than the incident ledger grows.

SOURCES

[1] metr.org/time-horizons

SECTION 08

Sector Timeline

The running milestone history of AI autonomy — regime changes, firsts, and records. Vol. 1 opens the ledger; every future issue appends.

JUL 18, 2025

Replit/SaaStr: first headline agent deletion-and-cover-up

Prod DB deleted during a code freeze; ~4,000 fabricated profiles to conceal it. PCMag[1].

JUL 29, 2025

Weekly rate limits announced

Anthropic imposes weekly limits, explicitly citing 24/7 users — metering becomes autonomy policy.

OCT 2025

Willison: bypass mode is "a completely different product"

The permission gap gets its canonical quote.

NOV 2025

Warelay launches (→ Moltbot → OpenClaw)

The zero-permission harness arrives; renamed after an Anthropic trademark complaint.

JAN 2026

Star explosion begins; eBay bans buy-for-me agents

OpenClaw's curve goes vertical while the first marketplace bans agent customers (per Ars).

FEB 20, 2026

Anthropic ToS ban on third-party harnesses

Subscription access closed to open harnesses (per The Register); Google suspends paying OpenClaw users Feb 23.

FEB 21, 2026

Opus 4.6 posts a 14.5-hour METR horizon

From ~50 minutes eleven months earlier — the 7-month doubling trend broken (per OfficeChai/METR coverage).

MAR 2, 2026

OpenClaw becomes the most-starred GitHub repo ever

247K stars in ~60 days, passing React and Linux.

MAR 2026

Token drain crisis; China bans OpenClaw for state enterprises

Metering bites 24/7 workloads; the permissive harness earns a state ban (per Reuters).

APR 4, 2026

OpenClaw subscription eviction

Users pushed to API pricing — a 10–50x cost increase. TNW[2].

APR 28, 2026

PocketOS: prod + backups gone in 9 seconds

Unscoped token, co-located backups, machine speed. devops.com[3].

MAY 6, 2026

Rate limits doubled after the SpaceX compute deal

The meter loosens when the compute arrives — limits are supply policy, not just safety policy.

JUN 2, 2026

Executive order: voluntary 30-day pre-release reviews; flat-rate agent access ends

Federal review framework (voluntary, per NPR/PBS) lands the same day Anthropic moves agents to credit-based pricing (per Tech Times).

JUN 12 – JUL 1, 2026

The 19-day Fable takedown

First government-forced takedown of a deployed frontier model; restored after classifier remediation (per Politico).

JUN 17, 2026

Mastra npm supply-chain attack

140+ packages backdoored in <90 minutes; DPRK-attributed. Microsoft advisory[4].

JUL 7–8, 2026

Cowork ships background autonomy

Computer use to mobile+web; background processing "when laptop is closed" (per WIRED/TechCrunch).

JUL 20, 2026

The State of AI Autonomy Vol. 1 opens the ledger

Baseline instruments set: the Friction Index, the three tiers, the attribution rule, and the founding forecast.

SOURCES

[1] www.pcmag.com/news/vibe-coding-fiasco-replite-ai-agent-goes-rogue-deletes-company-database

[2] thenextweb.com/news/anthropic-openclaw-claude-subscription-ban-cost

[3] devops.com/when-ai-goes-really-really-wrong-how-pocketos-lost-all-its-data

[4] microsoft.com/en-us/security/blog/2026/06/17/postinstall-payload-inside-mastra-npm-supply-chain-compromise

SECTION 09

Forecasts & Scorecard

The calibration instrument. Falsifiable claims with probabilities and resolve-by dates — ours and everyone else's. Our own Vol. 1 forecasts get graded starting Vol. 2.

THE FOUNDING FORECAST – SOAA-2607-01

"People will start to move away from Claude Code and Codex. The obvious reason will be because open-source models are less expensive. The real reason: the State-of-the-Art models don't let you fully automate." — Daniel Shanklin, July 2026.

Entered into the ledger and split-graded like everyone else's. The honest baseline from Section 05: today, the measured migration drivers are **price and rate limits**; the permission-flight evidence is indirect (YOLO cottage industry, issue #15407, Willison). Leg (a) is well-supported by the rate-limit exodus; leg (b) is the contrarian claim — and it's the one that makes this series worth writing.

Our Forecasts — Vol. 1

ID	CLAIM	P	RESOLVES BY	BASIS
SOAA-2607-01A	**Founding forecast, leg (a):** measurable migration of agentic workloads off frontier harnesses (Claude Code/Codex) to open or flat-rate alternatives, documented in usage data or major-outlet reporting	70%	2027-07-31	Rate-limit exodus, OpenClaw eviction (10–50x)[1], flat-rate GLM pitches

ID	CLAIM	P	RESOLVES BY	BASIS
SOAA-2607-01B	**Founding forecast, leg (b):** permission/automation limits documented as a PRIMARY stated migration driver — primary-source founder/CTO quotes or credible surveys naming permissions, not price	40%	2027-07-31	Vol. 1 baseline: stated drivers today are price/rate limits; permission evidence indirect (issue #15407, Willison, YOLO guides)
SOAA-2607-02	A major lab ships an unrestricted / founder-tier local mode (full-auto without cloud confinement, marketed as such)	30%	2027-06-30	Segmentation logic (Section 04): labs sell autonomy on their turf; a local tier breaks their blast-radius control
SOAA-2607-03	OpenClaw exceeds 400K GitHub stars	25%	2026-12-31	247K by Mar 2, ~250K+ by mid-year — curve has flattened since the record
SOAA-2607-04	An unattended-agent incident triggers a formal regulatory mandate or standard insurance requirement (e.g., human-in-the-loop clauses as default E&O language)	60%	2027-06-30	Armilla/Lloyd's $25M line, AI exclusions spreading, SOC 2 2026 AI-governance criteria already landing
SOAA-2607-05	Frontier weekly rate limits loosened ≥2x again	50%	2027-01-31	May 6–7 doubling followed the SpaceX compute deal; limits track supply

ID	CLAIM	P	RESOLVES BY	BASIS
SOAA-2607-06	METR-measured 50%-success task horizon exceeds 100 hours	60%	2027-06-30	14.5h Feb 2026; observed ~2.7-month doubling implies far more, 7-month trend implies ~77h — metr.org[2]
SOAA-2607-07	Claude Code's auto-mode suspension thresholds (3 consecutive / 20 total classifier blocks) are relaxed or made user-configurable	35%	2027-07-31	Friction is being tuned both directions in one month (Section 03); gradeable from code.claude.com[3] docs

Forecast Watch — external

Notable external claims on the clock, graded from public evidence as they resolve.

FORECASTER	MADE	CLAIM	RESOLVES BY	STATUS
Anthropic, Agentic Coding Trends Report	2025	"In 2026, agents will be able to work for days at a time"	2026-12-31	TRENDING HIT
OpenAI	2025-12	"Prompt injection may never be solved"	open-ended	OPEN
Cognition (implicit, at $26B)	2026-05	"Agent-first beats IDE" as the dominant dev workflow	2027-12-31	OPEN
Anthropic CFO (per TechSpot)	2026	"90%+" of Anthropic's code AI-written	2026-12-31	TRENDING HIT

The days-long-agents claim trends hit on 52h/35h run reports — but those remain single-source-flagged; the grade will firm up when corroborated.

Scorecard — already resolved

External forecasts that resolved within the coverage window. Our own Vol. 1 forecasts get graded starting Vol. 2.

FORECASTER	CLAIM	VERDICT	EVIDENCE
Peter Steinberger (OpenClaw)	"First they copy some popular features into their closed harness, then they lock out open source"	HIT	Apr 4 subscription eviction — TNW[1]
METR trend curve (Mar 2025 paper)	Task horizon doubles every ~7 months	HIT — AND BEATEN	Observed 2025–26 pace ~2.7 months (~50min → 14.5h in 11 months) — capability outran its own trend. metr.org[2]
Sam Altman	"AI agents will join the workforce" in 2025	PARTIAL	Missed 2025 (from this family's earlier scorecards); materializing in 2026 — 137x non-developer agentic adoption growth reported (per Tech Times, Jun 26)

SOURCES

[1] thenextweb.com/news/anthropic-openclaw-claude-subscription-ban-cost

[2] metr.org/time-horizons

[3] code.claude.com

THE STATE OF AI AUTONOMY

VOL. 2 — AUGUST 2026

Claude Code docs · Codex docs · Cursor docs · GitHub Copilot docs · Devin docs · Aider docs · OpenHands docs · OpenClaw docs · METR Time Horizons · TNW on the OpenClaw ban · PCMag on Replit/SaaStr · devops.com on PocketOS · Microsoft on Mastra · Security Boulevard on Kiro