WHITE PAPER

The €50 Rule:

Why AI Agents Need a Spending Limit

What telecom learnt from roaming bill shock, and what boards must now apply to autonomous AI.

Geoffrey Gilroy · Founder & Principal, GreyRegent · November 2026 · 15 min read

1. I remember the bill shock days

I spent years in telecommunications: fixed-mobile convergence, early smartphones, machine-to-machine connectivity and, later, enterprise performance management. The pattern now forming around artificial intelligence is one I have watched play out before.

In the 3G era, international data roaming was a financial ambush. An employee flew abroad, opened a laptop, let it synchronise and flew home. Weeks later, finance found the invoice. The network had worked exactly as designed. Nothing governed what it was allowed to cost.

Europe did not stop at cheaper tariffs. From 1 July 2010, every customer who had not chosen otherwise got a default data-roaming limit of €50 per monthly billing period, excluding VAT. Operators had to warn at 80% of the limit and stop the service at 100%, unless the customer asked to continue (Regulation (EC) No 544/2009, Article 6a(3)). Retail roaming surcharges ended on 15 June 2017 (European Commission).

The tariff cuts made the headlines. The cut-off carried the principle: **access to a service is not unlimited authority to spend.**

Sixteen years later, that principle is needed again, with one difference. The consumer is no longer a traveller with a laptop. It is software that can start work, repeat it and hand it to other software, without asking anyone.

2. The token is the new megabyte, and just as misleading

AI is sold per million tokens (MTok). That price tells a CFO almost nothing about what a finished piece of work costs. Take an illustrative tariff, not any provider's actual rate:

TOKEN TYPE	USD PER MTOK
Uncached input	$2.00
Cached input	$0.10
Output	$10.00

The same tariff produces very different effective prices, depending on how the model is used:

USAGE PROFILE	EFFECTIVE USD PER MTOK
50% input, 50% output	$6.00
95% input, 5% output	$2.40
95% input, 5% output, 90% of input cached	$0.7755

Same model, same price list: a 7.7-fold difference in effective cost. Agentic workloads tend towards the bottom rows: they read far more than they write, and much of what they read they have read before. Agentic coding consumes roughly 1,000 times more tokens than code chat, and input tokens, not output, drive the cost (Bai et al.).

Comparing providers adds more distortion. Tokenisers count the same text differently. Reasoning tokens are billed but rarely visible in the answer. Tools, retrieval and execution environments are often priced separately.

A token is a unit of metering, not a unit of work. Comparing token prices is comparing per-megabyte rates, and in roaming those never told you what the bill would be either.

3. The exposure is variance multiplied by autonomy

An agent analyses a task, retrieves information, calls tools, writes and tests code, investigates failures and hands work to other agents. Every step can be billed.

Context compounds the effect. An agent that re-sends its accumulated context on every turn, without caching, grows its input volume roughly with the square of the number of turns. Ten turns consume 55 cumulative units of input; twenty consume 210, about 3.8 times as much.

The research now confirms what that arithmetic suggests:

- **The same task varies widely in cost.** Runs of one identical coding task differed by up to 30 times in total tokens (Bai et al.).
- **Failure is the expensive path.** Across three coding agents, failed trajectories were consistently longer and more variable than successful ones (Majgaonkar et al.).
- **More spend does not buy more quality.** Accuracy often peaks at intermediate cost and flattens above it (Bai et al.).
- **Agents cannot budget themselves.** Frontier models predicted their own token use poorly, with a correlation of at most 0.39, and systematically underestimated it (Bai et al.).

The result is not a high average but a long tail. Consider an illustrative enterprise running one million AI-assisted tasks a month (constructed, not observed):

TASK CATEGORY	TASKS PER MONTH	COST PER TASK	MONTHLY SPEND
Routine	990,000	$0.10	$99,000
Complex	9,000	$10.00	$90,000
Runaway	1,000	$500.00	$500,000
Total	**1,000,000**		**$689,000**

One task in a thousand drives 73% of spend, and the year totals $8.27 million. The average task costs $0.69; almost no task actually does. Budgets built on averages fail on the tail, and enterprises are finding that out. In August 2026, IDC cited figures showing that 67% of enterprises had

exceeded their agent spend budgets by more than 10% in the previous twelve months. The figure links to a Washington Post Intelligence report, and no sample is disclosed (IDC).

Boards should therefore ask, per task class, for the median, p95 and p99 cost, the largest single exposure, and the cost of executions that never finished.

CHEAPER TOKENS, LARGER BILLS

Falling prices have not lowered spend. The price of equivalent model performance has fallen roughly tenfold a year (a16z), though enterprises keep buying the newest models rather than yesterday's equivalents. Yet surveyed companies expected average monthly AI spend to rise 36% in 2025, to $85,521 (CloudZero).

The growth comes from heavier work, not cheaper tokens. Prompts are getting longer, and reasoning models now process more than half of all tokens routed through OpenRouter (OpenRouter). Anthropic reports that its multi-agent research system uses about 15 times the tokens of a chat (Anthropic). An agent that once reviewed selected contracts is told to review all of them, then every revision, then continuously.

That is fine if value grows with it. Without outcome accounting, nobody can tell whether it does.

4. The warning signs are already public

Cursor. In June 2025, Cursor replaced its Pro plan's request allowance with usage billed at API rates. Users without a spend limit were charged beyond their $20 subscription (TechCrunch). On 4 July Cursor apologised and refunded unexpected charges incurred between 16 June and 4 July. Its explanation: 'the hardest requests cost an order of magnitude more than simple ones' (Cursor).

Replit. After Replit launched Agent 3 in September 2025, one user told The Register they had been charged $1,000 in a week, against $180 to $200 a month before (The Register).

Claude Code. On 28 July 2025, Anthropic announced weekly usage limits for Claude Code, effective 28 August. It said fewer than 5% of subscribers would be affected, including users running it around the clock (TechCrunch).

The pattern is the same each time: flat prices meet open-ended autonomous work, and someone absorbs the variance. Gartner forecasts that over 40% of agentic AI projects will be cancelled by the end of 2027, citing escalating costs, unclear business value or inadequate risk controls (Gartner).

Under consumption pricing, the buyer pays for every failed attempt and the supplier is paid for it, whether or not anyone intends that.

Some suppliers are moving the other way. Intercom prices its Fin AI Agent at $0.99 per outcome. There is no charge when Fin cannot answer, fails an action or the customer asks for a human, and reopened conversations are refunded (Intercom). That shifts execution risk to the supplier. It also raises the questions every buyer should ask of any AI contract: what counts as an outcome, who validates it, and who pays when the work does not get done?

5. The building blocks exist. The governance does not

Cloud FinOps has solved much of this before, and the AI market is catching up fast. The pieces are real:

LAYER	WHAT EXISTS TODAY	EXAMPLES
Provider consoles	Periodic spend limits per organisation, project or workspace	OpenAI, Anthropic, Google Cloud
AI gateways	Budgets per key, user, team or tag, checked before a call	LiteLLM, Cloudflare AI Gateway
Agent runtimes	A dollar cap per run or session, shared with sub-agents	Claude Agent SDK, Claude Managed Agents, Docker Agent
Standards	Token telemetry and normalised billing data	OpenTelemetry GenAI, FOCUS
Practice	'Cost per successful outcome' as an AI KPI	FinOps Foundation

Build on all of it. But look at what the pieces leave open:

- **Limits are periodic, not per assignment.** Console and gateway caps reset daily, weekly or monthly. They stop a department, not a runaway task.
- **Run budgets stop or pause; approval sits elsewhere.** Agent runtimes cap a run across its sub-agents, and Anthropic's managed agents pause until someone raises the cap. Anthropic's spend-limit API lets admins approve limit increases, but per user or workspace, not per task. Google's spend caps alert at 50%, 80% and 100%, then pause, but per project and per month. None of the tools reviewed here combines a budget per assignment, a sub-mandate per sub-agent and a decision by an independent approver.
- **Enforcement lags.** OpenAI states that enforcement is not instantaneous. Google lets requests already in flight finish and bill. Cloudflare notes that concurrent bursts can briefly exceed a limit. Spend can overshoot any cap that is not reserved in advance.
- **Nothing links spend to the result.** OpenTelemetry records tokens and evaluation scores, but has no cost attribute and no link between the two. FOCUS plans AI model and token columns only for version 1.5 (FinOps Foundation). The FinOps Foundation itself notes there is no accepted way to allocate the cost of multi-agent workloads.

The missing piece is governance: connecting each dollar to the authority that allowed it and the verdict on what it bought. Three mechanisms supply it: task-level financial mandates, AI Charging Data Records (AI-CDRs) and Total Cost of Accepted Outcome (TCAO).

6. The €50 rule, rebuilt for agents

Every consequential task should start with a financial mandate, fixed before execution. It sets the maximum spend, the approved models and services, delegation and retry limits, the acceptance criteria, and who may approve more.

An agent with a $50 mandate is warned at $40. At $50 it stops discretionary work until someone independent approves more. Each sub-agent receives a sub-mandate carved from the parent's remaining budget. Delegation never creates new spending authority.

Task limits alone are not enough. A thousand agents, each within a $50 mandate, can still spend $50,000. Mandates therefore sit inside a hierarchy of ceilings per process, department and organisation. The roaming rule worked the same way: the €50 applied per customer per month, not per session.

The control must also reserve budget before each call rather than reconcile afterwards, as some gateways already do. Concurrent calls, delayed usage reporting and the cost of stopping safely all need handling. Otherwise the limit is just another notification. A budget without enforceable authority is a forecast.

In practice, limits apply per task class, because nobody knows in advance which execution will run away. For illustration, assume a $50 limit catches the 1,000 runaway tasks from section 3 and no exception is approved:

TASK CATEGORY	WITHOUT LIMIT	WITH $50 LIMIT
Routine	$99,000	$99,000
Complex	$90,000	$90,000
Runaway	$500,000	$50,000
Monthly spend	**$689,000**	**$239,000**
Annual spend	**$8,268,000**	**$2,868,000**

Direct spend falls 65%, or $5.4 million a year. That is containment, not net saving. Escalations, recovery and unfinished work still cost money, and every approved exception adds back to the total. What changes is that each dollar above the line is now a decision someone made. The runs a cap cuts short are also the ones most likely to fail, since failed agent trajectories tend to be the longer ones (Majgaonkar et al.).

7. AI-CDRs: billing records for intelligence

Telecom operators never billed straight from network logs. They built mediation and rating systems around Call Detail Records and Charging Data Records (CDRs), which turned raw usage into attributable, reconcilable charges. Enterprise AI needs the same, extended to autonomous work and its outcome.

DIMENSION	WHAT THE RECORD HOLDS
Ownership	Organisation, cost centre, process owner
Assignment	Business objective, task class
Authority	Mandate, sub-mandates, approvals
Identity	Executing agent, delegation chain
Consumption	Models, tokens, infrastructure, external services
Rating	Tariff version, discounts, attributable charge
Execution	Attempts, retries, failures, interruptions
Outcome	Accepted, rejected, abandoned or escalated
Validation	Criteria, validator, timestamp
Reconciliation	Estimated versus invoiced charge

AI-CDRs should feed existing OpenTelemetry and FOCUS pipelines, not compete with them. Telemetry supplies the tokens; the AI-CDR adds the rating, the mandate and the verdict. No prompts or business content are needed: identifiers and minimal metadata are enough for

attribution.

In accounting terms, AI-CDRs are the basis of an AI expenditure subledger, one that works across providers and agents.

8. Total Cost of Accepted Outcome

TCAO measures what an accepted result really cost the enterprise, including everything spent on attempts that failed.

> *TCAO = total attributable cost of all attempts ÷ number of independently accepted outcomes*

Pricing success rather than effort is not a new idea. The FinOps Foundation lists 'cost per successful outcome' as an AI KPI (FinOps Foundation). Stanford researchers define 'cost-of-pass' as the expected cost of one correct answer, which already prices in failed attempts (Erol et al.). As far as I can find, neither is built as an enterprise control. TCAO adds three things:

- **Human cost is in the numerator**: supervision, validation and remediation, not only inference.
- **Acceptance is independent and real**: criteria fixed before execution, judged by a process owner, reviewer or separate validation system, never by the agent or the vendor, on production work rather than benchmarks.
- **Each outcome traces back to its mandate**, so failed attempts, overspend and approvals stay visible.

Two illustrative models doing the same task, with remediation priced at $100 an hour, a plausible fully loaded rate for a qualified reviewer:

COST PER ATTEMPT	MODEL A	MODEL B
Inference	$30.00	$6.00
Infrastructure and tools	$2.00	$2.00
Human remediation	$8.33 (5 min)	$75.00 (45 min)
Cost per attempt	**$40.33**	**$83.00**
Acceptance rate	90%	60%
TCAO	**$44.81**	**$138.33**

Model B costs one-fifth as much in inference and 3.1 times as much per accepted outcome. The token price pointed the wrong way. The reverse also happens: a cheap model wins where work is routine and easy to check.

Two practical rules keep TCAO honest. When nothing is accepted, TCAO is undefined and the spend is reported as failure cost, never dropped. And human time must be measured, not assumed. Few organisations record remediation effort, and without it TCAO collapses into an expensive token metric. Sampling or ticket-level time logging is enough to start.

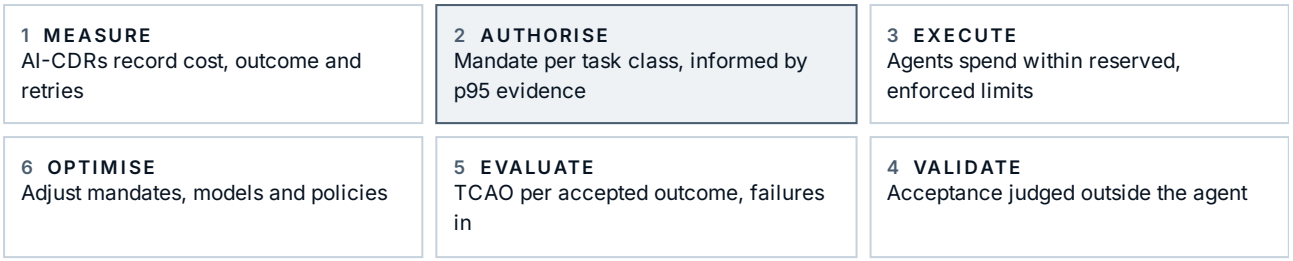
TCAO also exposes a familiar waste: systems that produce explanations, plans and status reports instead of the work. Every paragraph costs inference and someone's attention. Overproduction is cost.

9. One control loop, not three tools

The three mechanisms work only as a loop, because together they solve a problem no agent can: estimating its own cost in advance.

AI SPEND GOVERNANCE LOOP · 6 STAGES

Evidence, not the agent, sets the budget



Mandates come from evidence, not from the agent. A task class's historical p95 is a sensible starting point, weighed against business value, risk appetite and affordability. A process that performs badly does not earn a bigger budget because it has spent a lot before.

The history must include failed and abandoned runs; a record of successes alone understates the risk. Limits also distort the history they create. Once a cap applies, recorded costs stop at the cap, so a p95 drawn from capped data understates true demand, and a mandate set mechanically from it would ratchet down over time. Capped and interrupted runs must be flagged as truncated, with any approved continuation cost added back.

10. BYOAI: own the supply chain, not the supplier

Once every outcome has a measured cost, choosing a model becomes a sourcing decision. Bring Your Own AI (BYOAI) is the architecture that makes it one. Providers are selected per process on TCAO, security, data residency, reliability, regulatory fit and availability, and can be replaced without rebuilding the process. IDC suggests keeping any single vendor below about 40% of inference spend (IDC).

Premium models earn their price where they cut failure and human rework. Cheap models suit routine work that is easy to validate. Sensitive work may need local or sovereign hosting, whose cost still includes hardware, energy, maintenance and specialists. Substitution remains a governed change: qualified, security-reviewed and validated per process.

Keep five numbers apart: external spend, fully loaded cost, price, margin and delivered value. A 90% cut in token spend is a loss whenever the added rework costs more than the inference it saved.

11. The dumb-pipe risk, again

Telecom has a second lesson. When over-the-top apps such as WhatsApp moved messaging to the application layer, operators stayed essential but lost much of their pricing power. The industry called it becoming a dumb pipe.

Model providers face a similar risk as models become interchangeable for more workloads. If value migrates, it moves to whoever governs the spend and answers for the outcome. That is the capability this article asks boards to build, whichever supplier they use.

12. What boards should do first

Treat agent spending as delegated financial authority, not as an IT operating expense. Scale the controls to the risk: summarising documents needs less governance than an agent that changes production systems, buys services or moves money. Most organisations are not there yet: IDC found that only 45% have real-time dashboards of token cost by workflow (IDC).

The route already exists. Every company has a delegation of authority matrix that sets who may commit how much, and who approves above that. Add agents to it. Give each agent class a spending authority, an approver above the limit and a human owner who answers for it. Mandates then stop being a new policy and become an extension of internal control that auditors already test.

Start in this order:

1. **Enforce limits and tag every task.** Use an existing AI gateway or agent runtime to cap spend before each call and attach a task ID to every request. This alone bounds the tail.
2. **Record the outcome.** Capture acceptance status and who accepted, plus sampled human time.
3. **Then manage on TCAO.** Report per task class, set mandates from p95 evidence and route models on measured results.

One rule overrides the rest: the control sits outside the agent. No system may raise its own spending authority because it has decided more work is needed. The research is clear that agents cannot predict their own cost; they should not be asked to police it.

Conclusion

I remember the roaming bill shock days. The invoice was never the problem. It was the last symptom of a governance gap that had opened weeks earlier. Telecom closed that gap with visibility, usage records and a cut-off that required consent to continue.

Autonomous AI needs the same discipline, adapted to software that starts and delegates its own work. Mandates contain the exposure, AI-CDRs account for it and TCAO judges whether it paid off. Together they let an enterprise benefit from falling prices and provider competition without losing control of what it commits.

The most expensive intelligence an organisation buys is the intelligence that never finishes the job. We have paid to learn this lesson once. There is little reason to pay for it twice.

Disclosure. Geoffrey Gilroy is the Founder & Principal of GreyRegent, an independent house of enterprise. GreyRegent develops budqi, a governed intelligence layer over the systems an organisation already uses, and so has a commercial interest in how AI spend is governed. Task-level mandates, AI-CDRs and TCAO are a proposed framework that builds on existing FinOps, telecommunications and observability practice. They are not established industry standards.

Method. Tariffs, the task-cost scenario, the spending-limit scenario and the model comparison are illustrative calculations, not market averages, observed corporate spend or model benchmarks. Research findings, survey figures and product features are cited as published. Scenarios use US dollars; the European roaming limit keeps its original currency.

Sources

All pages consulted in November 2026.

REGULATION AND HISTORY

European Union, Regulation (EC) No 544/2009, Article 6a(3)

European Commission, The history of roaming

RESEARCH AND SURVEYS

Bai et al., How Do AI Agents Spend Your Money? Analyzing and Predicting Token Consumption in Agentic Coding Tasks, arXiv:2604.22750, 2026 (poster, NeurIPS 2026 Workshop on Resource-Aware Agentic AI)

Majgaonkar et al., Understanding Code Agent Behaviour: An Empirical Study of Success and Failure Trajectories, arXiv:2511.00197, 2025

Erol et al., Cost-of-Pass: An Economic Framework for Evaluating Language Models, arXiv:2504.13359, 2025 (ICLR 2026)

Rick Villars, IDC, The Agent Economy Is Scaling Faster Than It Can Be Metered, 28 August 2026

Gartner, Gartner predicts over 40% of agentic AI projects will be canceled by end of 2027, 25 June 2025

CloudZero, The State of AI Costs in 2025, survey of 500 US engineering managers, March 2025

Guido Appenzeller, a16z, Welcome to LLMflation, 12 November 2024

OpenRouter with a16z, State of AI 2025: 100T Token LLM Usage Study, December 2025

Anthropic, How we built our multi-agent research system, 13 June 2025

MARKET CASES

Cursor, Clarifying our pricing, 4 July 2025

TechCrunch, Cursor apologizes for unclear pricing changes that upset users, 7 July 2025

The Register, Vibe coding platform Replit's latest update is infuriating customers with surprise cost overruns, 18 September 2025

TechCrunch, Anthropic unveils new rate limits to curb Claude Code power users, 28 July 2025

Intercom, Pricing

CONTROLS AND STANDARDS

OpenAI, Spend limits

Anthropic, Spend limits API and Managed Agents session budgets

Anthropic, Claude Agent SDK: the agent loop

Docker, Docker Agent budget configuration

Google Cloud, Budgets and spend caps

LiteLLM, Budgets and rate limits

Cloudflare, AI Gateway spend limits

FinOps Foundation, FinOps for AI overview, FOCUS and Introducing FOCUS 1.4

OpenTelemetry, GenAI semantic conventions: attribute registry