The *moving* boundary

A change control problem that will define the next leaders

Chris Thorpe,
Agentic AI CTO, Pegasystems
July 2026

PEGA

1. The boundary is moving

Agentic AI is not a faster chatbot. It's a different economic system – and the market is pricing it wrong.

Token pricing was built for one problem space and is being stretched across two profoundly different ones. The gap is invisible in a demo. And at production volumes, it shows up directly in the operating budget.

Two distinct problem spaces are being sold under one label.

- The **first** - coaches, copilots, AI-adjacent tools, coding assistants — is billed per seat, bounded by the human in the loop, and is where most of today's enterprise AI spend lives. It works because the human is the rate-limiter on consumption.

- The **second** is agentic operations: autonomous workflows at production volumes, where consumption is bounded by the agent's reasoning chain – and by the supplier's tokenizer and pricing model. This market is in its infancy. Gartner's latest research puts the share of organizations that have deployed agents in production at **17%**; Capgemini puts at-scale deployment at just **2%**.[1]

- Gartner also forecasts agent software spending will roughly double in a single year – from **$206 billion in 2026 to $376 billion in 2027** – and warns that more than **40%** of agentic projects will be cancelled by the end of 2027, citing escalating costs, unclear business value, and inadequate risk controls.[2]

The market will dwarf the chatbot era it's currently being lumped in with. And a meaningful share of the projects funded against it will fail on unit economics before they ever reach production

This paper does the math. Then it lays out the operating model that puts unit economics back in the bank's control.

[1] Gartner, Worldwide AI Spending Forecast, May 2026. Capgemini Research Institute, "Rise of agentic AI: How trust is the key to human-AI collaboration", 2026.

[2] Gartner, ibid.

2. The deterministic majority

Most banking processes are deterministic – not by accident, but because regulation, audit, and decades of operating discipline made it so.

Reasoning is required only at specific junctures: synthesizing evidence on a marginal case, framing a reply, deciding what to escalate. Everything else is structure.

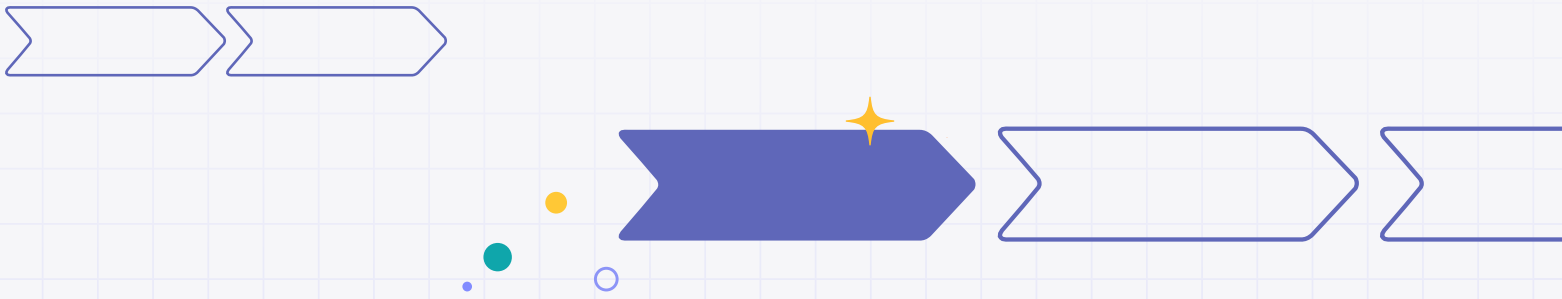
What's changed is the line.

Until recently, what could be automated was bounded by what classical code could express. Anything requiring judgement, language, or pattern recognition stayed manual. Agents have moved that line.

The banks that win this cycle will use agents to redesign entire process areas – complaints, call centers, the back office – domains that were built when judgement was beyond automation. Not the ones running thousands of agents on processes that were never reconsidered. Fewer than one in five enterprises today report the data and technology maturity to do this work well.[3] This is a change control problem more than a technology problem.

Even with the right architecture, operating problems follow. Agents drift: Their behavior today is not their behavior next month. Audit defensibility gets harder – because you have to reconstruct not just what the agent did but why, in language a regulator will accept. Most agentic vendors describe these problems in slides. Most banks discover them in production.

Agentic AI adds contained capability on top of deterministic process – while shifting the line of what counts as process in the first place. That boundary is the most important architectural decision your team will make in 2026 and beyond.



3. The math at production volumes

Let's make this concrete.

Loan exception adjudication is the work that begins when a mortgage, auto, or consumer-credit application falls out of the rules-based underwriting path. A top-three U.S. originator processes several million credit decisions a year – with 5–15% falling to exception handling. The queue across all consumer credit: roughly 300,000 cases per year.

Most of the work is deterministic: structured intake, rule-defined triggers, regulatory reason codes, adverse-action notices, audit trail. Reasoning is required at two junctures: synthesizing evidence into an underwriting judgement on the marginal case and framing the adverse-action notice on decline.

The copilot version: A human underwriter handles the case; the copilot assists with document review, income reconciliation, precedent retrieval, and a draft notice. A defensible estimate: 75,000 tokens per case.[4] At a blended frontier-model rate of $9 per million tokens,[5] on 300,000 cases per year, annual cost is approximately $200,000. The labor displaced more than pays for the spend.

The agentic version: Autonomous underwriting requires ingesting the full application, document-by-document analysis, precedent reasoning, an underwriting memo, fair-lending self-review, and the regulatory audit trail. Per-case token consumption falls in a 2–7 million range. Using 3 million tokens for the typical case on 300,000 cases per year at $9 blended: $8.1 million baseline. At the 7-million-token end, the same volume runs to roughly $19 million.

The driver is compounding, not linear. An agent without a planning layer re-reasons its own process at every step. Token consumption grows quadratically with process length: a ten-step workflow costs closer to 55 times a one-step call.

The harness version: The deterministic majority run as process, and reasoning is confined to the two critical junctures. Per-case consumption drops by an order of magnitude. The workflow lands far closer to the copilot's $200,000 than to the agent's $8 million – for the same outcome, on an audit trail the bank owns.

[3] Token estimates synthesised from published agentic benchmarks (τ-bench, AgentBench) and observed production traces; figures conservative.

[4] $9 per M tokens is a representative frontier-tier rate. It approximates Anthropic Sonnet 4.6 list ($3 input / $15 output per M) at a 50/50 input/output ratio, or Opus 4.7 ($5/$25) at an 80/20 split, the heavy-context shape typical of agentic underwriting. Comparable rates apply across major suppliers. A tier-one bank purchasing at scale will pay materially less than rack rate; the figure here functions as an upper bound on the worked example.

4. The operating harness

§2 argued that the work is redesign: reconsidering the process before agents are deployed against it. This section is the runtime half — the operating harness that turns that design into production behaviour you can govern, observe, and afford. Three elements have to be present.

1

The **first** is a planning function that runs before the reasoning model does. The default vendor pitch is that the agent plans its own work, deciding the sequence at runtime from a goal and a set of tools. Ad-hoc runtime planning is the most expensive way to learn what the work was. The harness keeps planning separate. A planner, sometimes deterministic and sometimes a small model, decomposes the work into named steps before the reasoning model is called. The reasoning model then does only its assigned step; it doesn't decide what to do next. The planner also governs context: each step sees the evidence it needs and nothing else, which is a cost control and a quality control in equal measure. Token consumption is bounded because the chain is bounded. The quadratic curve breaks to linear: each step sees only what it needs, not the accumulated history of the workflow, and audit defensibility holds because the decision tree is reconstructible without replaying the model.

2

The **second** is model independence — and for a tier-one bank the point is not the gateway. Most already run poly-model estates with failover and retesting discipline. The agentic-specific problem is variability: the same step behaves differently across models and versions, and in an agent-only system that variability compounds across the whole reasoning chain. The harness pins each step, so behaviour is testable step by step rather than workflow by workflow — and the same step can move to a different model or supplier when it has to, because models underperform on specific tasks and models go down. A frontier API unavailable for six hours is six hours of production agentic work that did not happen. The commercial exposure is real as well: vendors will raise prices, change tokenisers, and deprecate models on their schedule.[6] Predictive response caching sits inside this layer, and deserves more than a passing mention: a cheap routing model decides when a request can be served from a prior response, and at production volumes it is one of the few levers that materially moves the numbers in §3.[7]

3

The **third** is audit, state, and observability built in rather than bolted on. State management across long-running agents, audit reconstructibility a regulator will accept, reasoning-chain observability at the step level: these are the price of admission to a regulated production environment. Most current agentic stacks do not meet this bar. That is a specification the bank requires.

[5] - Not hypothetical: Anthropic's Opus 4.7 (April 2026) shipped at an unchanged $5/$25 headline rate with a new tokeniser that can consume up to 35 per cent more tokens for the same input - an effective price rise of up to a third, on the vendor's schedule, with no line item to point at.

[6] - Small-model pricing tiers as of June 2026, e.g. Anthropic Haiku 4.5 at $1/$5 per M tokens, OpenAI nano-class, Google Flash, are roughly 5x cheaper than frontier tiers and well-suited to routing and reuse decisions.

5. The real game

The instinct most banks bring to agentic AI is the one that worked for chatbots: find a use case, prove it works, scale it. That instinct does not transfer. The work is a different category, with different economics and a different change management problem.

The architecture in this paper is what makes agentic AI work in production. But it does more than that. It is also how the bank manages change over time. It gives you the dial.

This is the part most banks miss. The change control problem of agentic AI is a gradient, not a single yes/no decision. How much autonomy do you grant the agent today? How much next quarter, once you have six months of operating data? Without a harness, you have no dial: deployed or not, agent reasoning over the whole workflow or none of it, the vendor's audit trail or yours. With a harness, every element becomes a control surface. The planning function lets you tighten or loosen what the agent is allowed to decide. Model independence lets you route sensitive cases to safer models. Audit and observability let you watch the dial move and intervene when it overshoots.

That is what change management for agentic AI looks like: a control framework that lets the bank turn the dial on its appetite for AI risk over time, on the cadence the business can absorb. The alternative is a single board decision followed by a five-year roll-out, and that approach is what produces the failure mode Gartner is already measuring.

Agents have moved the line of what can be automated. That is the real game: business reimagination. Whole operating areas that were built when language and judgement were beyond machines are now candidates for redesign. The harness is what lets you take that on without betting the bank on a single decision.

Chris Thorpe,
Agentic AI CTO, Pegasystems