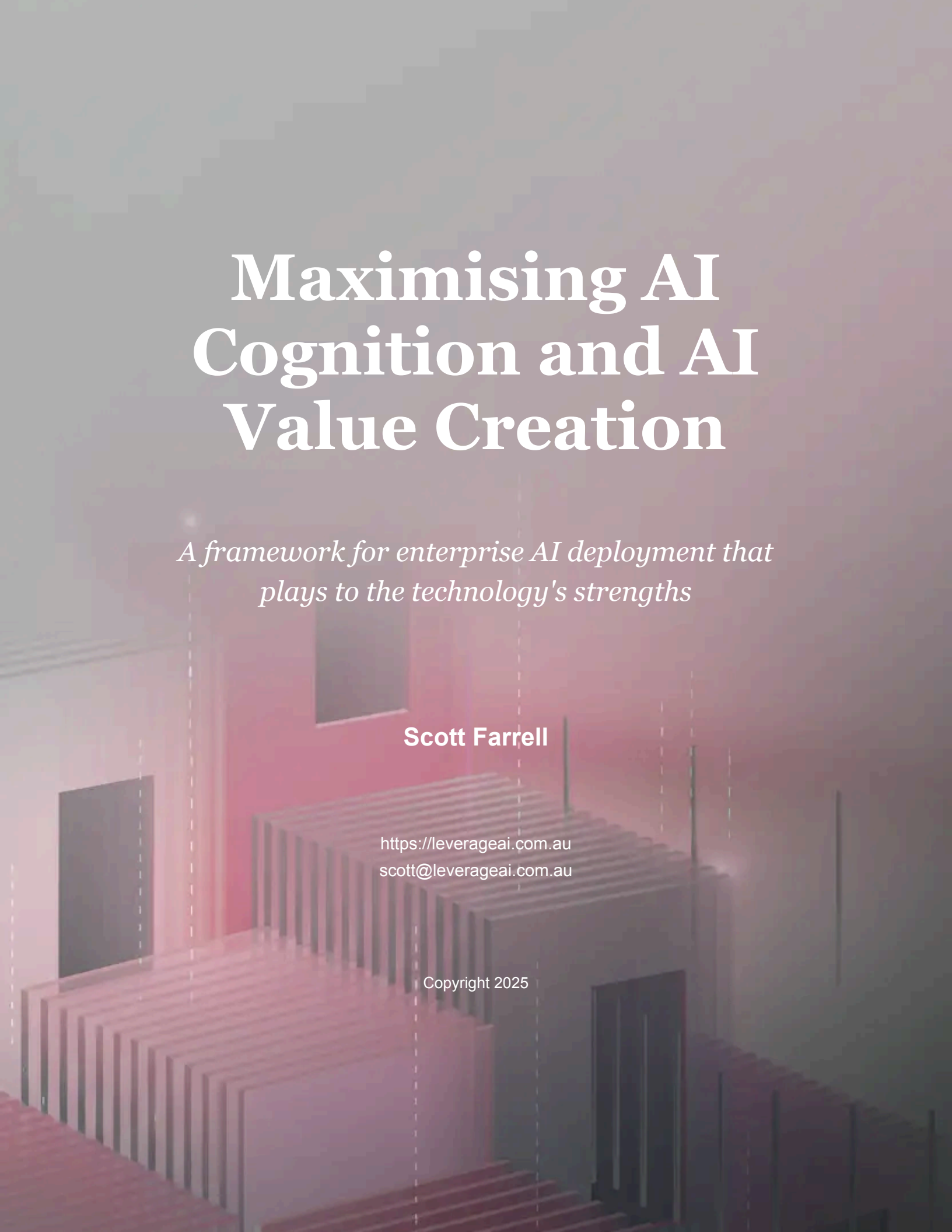

Maximising AI Cognition and AI Value Creation

A framework for enterprise AI deployment that plays to the technology's strengths

Scott Farrell

https://leverageai.com.au
scott@leverageai.com.au

THE UNCOMFORTABLE TRUTH ABOUT AI FAILURE

Setting the stage with failure statistics and the central question

95%

of corporate AI pilots show ZERO return on investment

MIT Media Lab, 2025

In the boardroom of a mid-sized insurance company, the mood was triumphant. After nine months and $2.3 million, their AI chatbot was ready to launch. Six weeks later, the same boardroom witnessed a different conversation. Customer complaints spiked. The chatbot was quietly pulled offline. The project became a cautionary tale.

This scenario isn't fictional. It's the pattern playing out across enterprises worldwide. And the scale should alarm every executive allocating capital to AI.

The Statistics That Should Alarm You

Despite $30-40 billion in enterprise investment, AI failure is officially the norm. Three-quarters of enterprises investing in AI see no meaningful return. They're running pilots, attending conferences, hiring consultants, buying tools—and getting nothing.

The gap between proof-of-concept and production has become a graveyard for AI ambitions. Only 26% of organizations have capabilities to move beyond POC to production. Just 6% qualify as "AI high performers" achieving 5%+ EBIT impact.

The AI Failure Landscape

70-85% of AI projects fail to meet expected outcomes

42% of companies abandoned most AI initiatives in 2025

2x AI projects fail at twice the rate of traditional tech projects

RAND Corporation; S&P Global

The Puzzling Contrast

Yet some companies achieve spectacular success. The contrast is stark: while 85% fail, 15% achieve returns of 3x to 10x on their investment. Klarna's AI assistant delivered a $40M annual benefit, replacing work of 700 employees. A single AI system delivering transformation, not incremental improvement.

The difference isn't the technology

Everyone has access to the same models. The difference is where and how they deploy.

The Question This Book Answers

What separates the 15% that succeed from the 85% that fail? It's not better AI models, bigger budgets, better vendors, or more expertise. The answer: successful organizations deploy AI where it has asymmetric advantage. They don't try to make AI work everywhere. They identify specific contexts where AI's strengths dramatically outweigh its weaknesses.

This book gives you the framework to make that distinction. You'll learn the real economic argument for AI, the latency vs. accuracy asymmetry that predicts failures, a deployment matrix showing where AI creates value, and the three versions of AI value—including Version 3 where transformation happens.

By the end, you won't just understand why 85% of AI projects fail. You'll have a systematic way to ensure yours doesn't join them.

THE REAL CARROT: COST OF COGNITION

Strip away the hype. What do companies actually want from AI? The answer isn't "robots"—it's something far more strategic.

If you ask a CEO why they're investing in AI, they'll rarely say the quiet part out loud. But when pressed, the answer almost always comes down to economics.

Here's the reframe that cuts through: firms buy AI because **thinking per hour** will be way cheaper than human thinking per hour. Let's give it a name: **cost per unit of useful cognition**.

100x

More thinking for the same spend

When marginal costs trend toward zero

The True Cost Comparison

For humans, that's not just salary. It's salary plus all the on-costs: management overhead, desk space, tooling, training, coordination overhead, and—let's be honest—meetings to decide what the meetings meant.

$150K–$180K

All-in cost of a $100K employee

— Before counting time in meetings that produce no decisions

For AI, you've got a different stack: model costs, platform infrastructure, integration, governance, monitoring, and incident response. So no, AI isn't "$1 per hour" when you account for everything.

But the proposition still holds: once the plumbing is in place, the **marginal cost** of an additional thinking task trends toward cents, not dollars.

> *"The carrot isn't AI. The carrot is: We can throw 100x more thinking at our problems than we used to, for roughly the same spend."*

How to Talk to Executives

Language matters. Talk about "robots" and executives hear science fiction. Talk about "automation" and HR starts resisting. But talk about **cheap cognition** and you're discussing **capability expansion**—shifting the conversation from cost centre to force multiplier.

What Lands

"We can apply 100x more thinking to our strategic problems. Cognition used to be our constraint. Now it's abundant. The question is: where do we deploy it for asymmetric advantage?"

When the Economics Work

The cognition cost advantage isn't universal. It kicks in when you have:

High Volume: Enough tasks that upfront investment pays off quickly.

Parallelisable Work: Tasks that run simultaneously without coordination overhead.

Time Flexibility: Work that doesn't need instant responses.

THE ASYMMETRY NOBODY TALKS ABOUT

Why AI thrives in batch processing but fails in live chat

72%

Say chatbots are a complete waste of time

— UJET survey, Forbes 2022

Two support teams receive the same customer question at 9am Monday. Team A deploys a live chat AI—customer expects response in 30 seconds, AI has one shot to be right. Team B uses an AI ticket system—response expected in 4 hours, AI can check history, cross-reference systems, escalate if uncertain.

Same technology. Same question. **Completely different success rates.**

The Latency-Accuracy Trade-Off

In live chat, AI needs to be fast, accurate, safe, and on-brand—all at once. That means stronger models, rich context, guardrails, verification steps. Each safeguard adds latency and cost.

Meanwhile, trained humans still win in low-latency, high-stakes, ambiguous situations. They bring tacit knowledge, read tone, pivot mid-conversation, make judgment calls. They don't need perfect accuracy on first try—they course-correct. "Let me check with billing" buys time without losing trust.

AI in live chat doesn't get these affordances. The data confirms it: 78% of chatbot users are forced to connect with a human after failure. 80% report increased frustration. 64% would prefer companies not use AI for service at all.

> *"You can't get it right on the third go around, because there's a real customer on the other end."*
>
> — The fundamental challenge

The Flip: When AI Has Time

Remove the time pressure and everything changes. With minutes or hours—not seconds—AI can read full history, cross-check systems, use multi-step reasoning, and escalate when uncertain. Customers already expect asynchronous responses.

Batch processing delivers 40-60% infrastructure cost savings vs real-time systems

Same volume. Same work. 80% fewer resources. Plus higher accuracy—broader view across datasets, time to verify patterns, lower false positive rates.

— Galileo AI; Zen van Riel, "Should I Use Real-Time or Batch Processing for AI"

Real-time requires 100 GPU instances running 24/7; batch uses 20 GPUs during off-peak hours. Infrastructure scales with work done, not peak capacity.

The Fundamental Asymmetry

AI's Weak Zone

Fast + accurate + one-shot

Live interactions with no error tolerance

Same technology, two fundamentally different games. AI is terrible at fast + accurate + one-shot. But brilliant at batch processing with time flexibility. Most companies deploy AI exactly where it's weakest—live interactions with no error tolerance—and wonder why it fails.

The Real Risk Isn't Hallucinations

Hallucination is one model making something up once. The risks that keep executives awake: **systemic wrongness** (AI faithfully executing a bad spec 24/7 at scale), **unobservable decisions** (no logs, no reasoning trails), **accountability blur** (bug in model, prompt, retrieval, or business logic?), and **silent drift** (behavior shifts slowly over months, no one notices until something hits regulators).

THE 2X2 THAT PREDICTS SUCCESS

A simple framework for predicting which AI projects will succeed

You've heard the pitch: AI will transform customer service, accelerate decision-making, automate operations. Then you deploy a chatbot that frustrates customers. Launch a real-time analytics system that costs three times more than projected. Build an autonomous approval engine that your compliance team refuses to certify.

The problem isn't the technology. It's that we lack a simple framework for predicting which AI projects will succeed versus which will fail.

The Two Dimensions That Matter

Every AI deployment decision can be mapped on two critical dimensions:

X-Axis: Latency Tolerance. Can this work wait? Not "should it wait" but "can it wait without breaking the business process or disappointing the customer?"

Left side requires responses in seconds—live chat, real-time fraud detection, synchronous API calls. Right side can wait minutes, hours, or overnight—ticket queues, batch analytics, document processing.

40-60%

Cost savings from batch processing vs real-time

Bottom-right quadrant advantage

Y-Axis: Consequence of Error. What happens if the AI gets it wrong? Bottom means cheap to fix—internal notes wrong, easy re-run, trivial customer inconvenience. Top means expensive—compliance violations, customer churn, legal exposure, safety incidents.

The Four Quadrants

> **Bottom-Right = Prime AI Territory:** High latency tolerance gives AI time to think. Low error cost means mistakes are cheap to fix. This is where the 40-60% cost savings live.

Bottom-right quadrant: This is where AI dominates. Ticket systems, overnight batch jobs, report generation. AI can take time to verify and iterate. Errors are cheap to catch.

Top-right quadrant: High-stakes but time-flexible. Deploy the "copilot plus gate" pattern. AI does deep analysis and drafts recommendations. Human reviews and approves before execution.

Left side (both quadrants): Human-led territory. When latency tolerance is low, human expertise wins. AI needs context and reasoning cycles to deliver quality. Real-time pressure forces shortcuts that degrade accuracy.

72% of consumers say

Chatbots are a waste of time

Forbes, UJET Survey

Common Misallocations

The most common AI project failures happen when organizations misread where their use case sits on the grid. The chatbot mistake: thinking live chat is bottom-right when it's actually left-side (low latency tolerance). Customers escalate to humans anyway. Frustration increases.

What actually works: Recognize live chat is left-side. AI surfaces context; human owns the conversation. Move ticket queues to the bottom-right where AI can resolve 40-60% autonomously with time to think.

Move the Conversation

THE THREE VERSIONS OF AI VALUE

From automation to amplification to frontier thinking—understanding the progression from cost reduction to entirely new capabilities

3

Versions of Value

Most stop at Version 1

Most AI conversations stop at "automation"—replace human tasks, save money, move on. But that's only Version 1, and ironically, it has the highest failure rate. To understand why some companies achieve 10x returns while 85% fail, you need to see the full progression.

Version 1: Same Work, Fewer People. This is the classic automation play. Replace a human task with AI—invoice processing, email triage, basic data entry. It's where most of the 70-85% failure rate lives because you're asking AI to beat humans in environments humans designed for themselves.

When does Version 1 work? Klarna's AI assistant replaced work of 700 customer service agents, delivering $35-38M annually. The key: they didn't deploy AI for instant-response live chat. They used it for ticket-based interactions where the system had time to read full history, cross-check policies, and escalate complex cases to humans.

Version 2: 10-100x More Thinking

This is where the conversation gets interesting. Instead of replacing people, you amplify cognitive output beyond what was economically feasible before. The fundamental shift: instead of *sampling*, you check *everything*.

Version 2 changes the economics: the marginal cost of additional thinking has inverted. One analyst sampling 50 transactions becomes AI checking all 500,000 daily transactions. Coverage increased 5,000x, false negatives dropped 60%. Companies achieving Version 2 report $3.70-$10.30 return per dollar invested.

Version 3: Previously Impossible Thinking. Now we reach the frontier—work that organizations don't even attempt today because it would be economically or politically insane. Large strategic initiatives take months with human teams not because the analysis requires months, but because of scheduling, coordination, and political navigation.

What becomes possible: Hyper sprints replacing months of committees with overnight AI exploration of thousands of strategic options. Marketplace-of-one personalization—per-customer offers, pricing, and service levels. Continuous strategic sensing of all customer interactions and market signals. Exhaustive scenario planning that would take a McKinsey team six months, delivered overnight.

Version 3: The Frontier

Work that was never feasible before because coordination overhead, calendar time, or organizational patience killed it. AI doesn't have calendar constraints, meeting fatigue, or political navigation requirements.

> *"If thinking was no longer our constraint, what would we attempt?"*

The uncomfortable reality: most organizations are stuck in Version 1 because it's easiest to imagine. They look at their org chart, identify tasks humans currently do, and ask "Can AI do this cheaper?" That framing guarantees you miss Version 3 entirely.

The Version 3 Question

What's one project we've never attempted because the coordination overhead was too high? What strategic analysis have we not done because assembling a 50-person team for six months was organizationally insane? Those are your Version 3 opportunities.

HYPER SPRINTS: REPLACING COMMITTEE-THINK

When thousands of possibilities are explored overnight, politics happen after seeing the full landscape—not during exploration

Picture the typical enterprise strategic project: ten cross-functional people, three months of meetings, and a PowerPoint deck that everyone can live with. The goal, whether stated or not, is rarely to find the **best** answer—it's to find an **acceptable** answer.

This isn't a criticism of the people involved. It's a structural constraint. Research on group decision-making reveals the mechanisms: when time is limited, less knowledge is shared. Decisions become negotiations between prior preferences rather than genuine exploration. Groups with similar compositions reach similar conclusions, often falling prey to groupthink.

> *"What's an acceptable answer that senior management will swallow? Based on experience of getting things past managers, not exploration of what's actually optimal."*
>
> — The Committee-Think Reality

The Hyper Sprint Alternative

200+

Iterations overnight

vs 3 weeks for human teams

What if you could replace that three-month committee process with something fundamentally different? Thousands of AI calls overnight exploring multiple frames, scenarios, and constraints. A full audit trail of what was considered and why. Human experts review in the morning and redirect the search based on insights. Politics happen *after* seeing the full landscape, not during exploration.

The crucial distinction: you're not asking AI to magically know the answer. You're asking it to systematically explore more possibilities than humans would have time for—like a chess engine exploring move trees. In chess, the human sets objectives and constraints. The engine explores the possibility space. The result? Move sequences humans would never have time to consider.

97%

Multi-model council accuracy vs 80% for single models

— Extended Thinking Performance Data

Extended Thinking: The Enabling Technology

This systematic exploration is only possible because of a fundamental shift in how AI systems work. Traditional models optimize for speed—they generate the first plausible answer. But newer systems like OpenAI's o1 and Claude 4.5's extended thinking mode work differently.

Instead of optimizing to answer instantly, we can give them compute budget to **think through the problem**. The AI explores multiple paths, evaluates trade-offs, considers edge cases—all before committing to an answer. On problems where a smaller base model attains somewhat non-trivial success rates, test-time compute can be used to outperform a 14x larger model.

From Committee-Think to Hyper Sprints

Committee-think optimizes for consensus under time pressure. Hyper sprints systematically explore thousands of options overnight, documenting reasoning and letting humans make the final call with full visibility. The same principle that revolutionized chess is now available for every strategic decision your organization makes.

MARKETPLACE OF ONE

Why we segment customers? Because treating each individually was too expensive. That constraint has changed.

$1T

Value shift from standardisation to personalisation

— McKinsey & Company

For decades, the economics of marketing and service delivery forced a fundamental compromise: segment customers into groups and design around the "average customer in segment X." It wasn't ideal—it was the only feasible approach.

The trade-off was predictable: individual preferences flattened to segment averages, outliers poorly served, one-size-fits-most becomes one-size-fits-none. Opportunities for personalised value creation left on the table.

The Economic Shift

What has changed is fundamental: the economics of personalisation have inverted. Previously, customisation was expensive—manual effort scaled linearly with customer count. Now, the cost structure has flipped.

McKinsey research quantifies the opportunity: personalisation typically drives 10–15% revenue lift, with company-specific results ranging from 5–25% depending on sector and execution capability. Shifting to top-quartile performance would generate over $1 trillion in value across US industries alone.

The New Reality

Recomputing per-customer recommendations now costs less than maintaining rigid segment-based rules that require constant exception handling.

62%
Higher engagement

80%
Better conversion

AI Magicx, 2025

> *"AI doesn't optimise for average—it adapts to context. Every interaction can be unique."*

What Becomes Possible

When the constraint of cognitive overhead disappears, entirely new design patterns become rational: per-customer campaign messaging, individualised timing and channel selection, personalised support flows, dynamic pricing based on individual behaviour, custom terms and conditions.

The Strategic Question

"If we had a smart assistant assigned to every customer and every employee, what would we design that we currently throw away because it's 'too complex to manage'?"

The answers to that question represent the marketplace-of-one opportunity. They're the products, services, and experiences you currently dismiss as "too complex to manage."

This Is Version 3

Marketplace of one represents Version 3 AI value creation—not doing old work faster (Version 1), not applying more thinking to existing problems (Version 2), but creating new work that was never feasible: a new class of product and service design.

AI AS COGNITIVE EXOSKELETON

The pattern that works across all three versions of AI value isn't about replacement. It's about amplification.

The difference between Version 1 and Version 3 thinking comes down to where you place AI in the workflow. Instead of asking AI to handle the customer interaction, ask it to prepare the human who will.

AI does the pre-work. Humans own the moment.

72% → 80%

Diagnostic sensitivity improvement with AI assistance

Medical professionals using AI-enhanced diagnostics significantly outperform either AI or humans alone

— EY, Human-Machine Economy

The Pre-Work Pattern

When a customer message arrives, AI mines CRM for context, infers what matters, pulls knowledge base articles, and surfaces a rich cockpit with suggested actions and risks to watch for. The human agent sees a prepared dashboard instead of scattered data sources.

90.2%

Multi-agent improvement over single-agent systems

Claude Opus 4.5 orchestrating specialized subagents

This gives the human three things: faster workflows with less clicking, more accurate decisions from richer context, and control over judgment and relationship handling.

Remember the latency-accuracy trade-off? If you surface information to the human rather than have AI make the decision, suddenly AI isn't making mistakes—it's providing options. Recovery from edge cases is instant because human judgment takes over.

Multi-Agent Orchestration

The augmentation pattern extends to teams of AI agents coordinating to amplify a single human's capability. A central orchestrator agent decomposes tasks and delegates to specialized worker agents with domain expertise—sales, legal, technical analysis.

> *"Multi-agent systems use about 15x more tokens than chats. For economic viability, they require tasks where the value is high enough to pay for increased performance."*
>
> — Anthropic, Token Economics Research

This is the trade-off: 15x token cost for 90.2% performance improvement. It makes economic sense for high-stakes decisions like M&A analysis or strategic planning. It doesn't make sense for routine email summaries.

Where This Pattern Applies

The cognitive exoskeleton pattern works anywhere humans face high-stakes, time-sensitive interactions that benefit from exhaustive preparation: sales calls with AI surfacing account history and competitor intel, support with AI preparing root cause analysis, medical consults with AI assembling differential diagnosis context, legal work with AI surfacing relevant precedent.

In all these cases, the interaction is high-stakes and time-sensitive. AI saturates the pre-work and side-work. The human owns the moment of judgment. Relationships and accountability stay with the human.

Not Replacement—Amplification

THE RIGHT QUESTIONS

From deployment failure to strategic transformation

70%

Of AI investment in people & process

Not technology alone

You've seen the deployment matrix. You understand the three versions of AI value. You know why chatbots fail and where batch processing wins. Now comes the implementation reality check.

Because "$1 per hour AI" is a fantasy. The question most organisations ask points them directly toward the 95% failure rate.

The Real AI On-Costs

Per-token pricing creates the illusion that AI is cheap. GPT-4.5 costs $10 per million output tokens. That sounds like pennies. But AI doesn't run on tokens alone—it requires an ecosystem to operate.

The Five Hidden Cost Categories

Infrastructure: Vector stores, orchestration, observability
Operations: Monitoring, alerts, anomaly detection
Governance: Policies, approvals, audit trails
Maintenance: Prompt tuning, version management
Change Management: Training, process updates, politics

Successful AI organisations don't just buy software—they build capabilities. That takes investment in infrastructure, process, and people. Organisations achieving value invest 70% of AI resources in people and processes, not just technology.

Early AI adopters report $3.70 in value per dollar invested, with top performers achieving $10.30 returns. But that takes patience. High performers commit 20%+ of digital budgets to AI, implement human oversight, and redesign workflows rather than bolting AI onto existing processes.

> *"Most organisations achieve satisfactory ROI within 2-4 years—much longer than typical 7-12 month software payback periods."*
>
> — Industry Analysis, 2025

The Three Questions

Most organisations start AI projects by asking the wrong question. Here's how the progression should actually work:

Wrong Question:

"Where can we put a chatbot?"

Technology-first thinking. Ignores deployment matrix. Leads to 95% failure rate.

Better Question:

"Where do we waste human thinking time on work that's slow, repetitive, or queued up?"

Identifies Version 2 opportunities (100x thinking applied). Finds batch/queue sweet spots. Focuses on proven ROI patterns.