

What's Changing in How You Deliver

IT Consulting

You have seen this on your delivery engine

"Your clients' procurement teams are showing up to renewals with AI-efficiency benchmarks that already estimated what AI-assisted delivery should cost. By the time your Head of Practice walks into the room, the rate has already been set."

01

Why is our margin drifting on work our rate card says should hold?

02

Does our delivery approach still mark us as the technology experts?

03

What happens to our senior bench if the junior training ground disappears?

More insights like this at grail.works

TECHNOLOGY DELIVERY WORKPLAN		
PHASE	FOCUS	OWNERSHIP
01 ARCHITECTURE & STRATEGY	• Business & technology alignment • Target architecture • Solution options & trade-offs • Roadmap & sequencing • Investment & value case	Partner Architect Principal
02 EXECUTION LAYER (AUTOMATED)	Automated build • Code generation • Configuration • Testing • Documentation	AI-Assisted Delivery
03 DELIVERY & INTEGRATION	• Solution build & integration • Data migration & cutover • Security & performance • User acceptance • Transition to BAU	Delivery Manager Team Lead
04 REVIEW & OUTCOMES	• Benefits realisation • Performance & adoption • Continuous improvement • Next horizon opportunities • Executive reporting	Engagement Director Partner

You have seen this on your delivery engine

Every Chief Operating Officer or Head of Delivery at a mid-sized European IT consulting firm has had the same version of Monday morning. Three items on your desk before 9am. First: notes from the Head of Practice on Friday's renewal conversation. The client's procurement lead asked why your quoted hours didn't reflect what AI-assisted delivery should cost. They had a benchmark. You did not expect it. Second: an exit interview summary. A senior developer left for a client-side role. Three years of knowledge about a specific legacy integration system, stored in their head and now gone. Third: the bench report. The junior cohort from 2024 is eighteen months in, and the promotion pipeline is stalling because the routine work that would have built their judgment is now running on agents.

You have run this delivery engine for ten years. The pattern used to be predictable. Hire the best consultants you could find. Hold the utilisation number. Protect the delivery quality. Grow the bench through repetition and mentoring. It worked.

Something changed. Your delivery margin is drifting on engagements your rate card says should hold. Your most senior consultants are the firm's most valuable asset and the most fragile. They carry client-specific judgment that has never been written down, and when they leave, the relationship goes with them. And you can name the six junior consultants whose development pathway you cannot currently explain.

This is not a you problem. It is an industry problem and a delivery-engine problem, happening at the same time. Most IT consulting COOs are treating them as two separate conversations. The ones who see them as one are going to own the next decade of European IT delivery.

This briefing covers what is happening in your sector, what it means inside your delivery engine, and the three questions your operations team should answer before the next board cycle.

IT consulting just lost the equation that ran it for a generation

Every technology wave your sector absorbed in the last twenty years gave clients a reason to pay for more consultant hours. Cloud migration expanded scope. Data strategy created new practice lines. DevOps transformed how delivery ran. Security compliance added mandatory expertise into every engagement. Each new wave made the rate card defensible because the work genuinely required the hours. The business has run on billable hours for a generation and it never once broke.

AI does not behave the same way. The client still needs the same architecture, the same integration, the same advisory on which system to replace and how to migrate. The scope is unchanged. But the routine analytical layer that absorbed most of the junior hours: the boilerplate code, the documentation, the standard testing, the basic integration work. All of it can now be done in a fraction of the time by tools the client already knows exist. Under hourly billing, faster delivery at the same rate card means fewer billed hours for the same engagement. Your sector knows this. Your pricing team knows this. But most delivery engines have not been restructured around it.

Your clients' procurement teams are showing up to renewals with AI-efficiency benchmarks that already estimated what AI-assisted delivery should cost. By the time your Head of Practice walks into the room, the rate has already been set.

The firms in your segment that are pulling ahead have done one specific thing. They stopped treating AI efficiency as a productivity gain under the existing model and started treating it as a pricing transformation opportunity. The firm that delivers the same engagement in fewer hours and charges a fixed project fee captures the efficiency as margin. The firm that keeps billing T&M watches clients demand the efficiency as discounts. Thirty percent of the former is margin. Zero percent of the latter is. That arithmetic is the strategic choice your sector is making right now.

Your delivery engine is feeling it every day

Your delivery engine. Project managers are working harder to hold delivery quality on engagements that were straightforward two years ago. The handoffs between business development, scoping, and delivery carry friction that was built when the function was organised around a 2018 consultant workflow and never redesigned. The tooling changed. The clients' expectations changed. The delivery architecture has not.

Your consultant bench. Your mid-level consultants built their judgment through repetitive technical work: integration templates, documentation cycles, testing runs, standard architecture reviews. That work is now largely agent-owned. They are more productive per hour. They also have fewer hours of judgment-building volume behind them than their counterparts from three years ago. Utilisation looks fine. The development pipeline is stalling in ways the utilisation number does not capture.

Your senior consultants. Your ten to twenty most experienced consultants carry between two and five major client relationships each. Those relationships were built on years of client-specific knowledge: the system quirks, the political dynamics, the procurement preferences, the things that are never in the project file. It has never been written down anywhere your next hire can access it. When a senior consultant leaves — for a client-side role, for a competitor, for retirement — the relationship is at risk before the transition conversation is finished.

Your knowledge continuity. Every engagement your firm runs produces a body of client-specific intelligence. The architecture decisions, the rejected alternatives, the reasons a particular approach worked in this specific context. Most of it lives in the project manager's notes and the senior consultant's memory. The firm that extracts that intelligence as a side-effect of delivering the engagement builds an asset. The firm that does not loses it every time someone new joins an account.

Your pricing and delivery model. Your current model was designed for T&M delivery when the margin was in the junior hours. The junior hours are the hours under most pressure. The model needs redesigning, and the COO is the only leader close enough to the delivery engine to see where it breaks and where it holds.

The firms pulling ahead are building one piece of infrastructure

Your sector is compressing from below because AI takes hours out of the junior billing model. Your delivery engine is compressing from above because your senior consultants are carrying client-specific knowledge as biography, not infrastructure. The same AI tools causing the first problem are the tools that took away the training ground that produced your seniors in the first place.

These are the same problem. The knowledge your firm runs on has never been separated from the person who holds it. Every margin problem, every continuity risk, every stalled bench is a symptom of that single unsolved thing. A junior with no routine work to do has no judgment to build. A senior who holds the relationship in their head creates a continuity risk the firm has been pretending is an HR question. A delivery engine priced on hourly input has no mechanism to capture the efficiency the tools are creating.

Look at the firms pulling ahead. They are not running three separate programmes. They are building one piece of infrastructure: the kind that makes every engagement a source of institutional knowledge, lets every consultant work at the level above where the agents now operate, and gives every client reason to stay because the relationship lives in the firm's systems, not in one person's memory.

You can tell which camp a firm is in from one question. Can you point to one engagement type in your delivery engine that is measurably faster, at higher quality, and at better margin this quarter than six months ago? Most IT consulting firms cannot. The ones that can have already made the move, and they are already winning the renewals you thought were yours.

You have eighteen to twenty-four months. The window is not about adoption. It is about which firms build the infrastructure first. The ones that do own the next decade of client relationships. The ones that wait inherit the clients who got used to better somewhere else.

What most firms are doing. What the winners are doing instead.

Every delivery COO at an IT consulting firm has had the AI conversation. Most are making the same four decisions. A smaller group is making different ones.

THE DEFAULT PATH	THE WINNERS' PATH
Hold T&M, push utilisation targets Same pricing model, harder targets. Clients demand efficiency discounts. Delivery margin drifts downward.	Migrate routine work to fixed-price delivery Price on the output, not the hours. The efficiency becomes margin, not a discount the client extracts.
Treat senior exits as a recruitment problem Hire replacements and watch the client relationship go to tender within six months of each departure.	Extract knowledge as a side-effect of every engagement Every project report, every architecture note builds institutional memory. The relationship stays when the consultant goes.
Deploy AI tools on top of the existing delivery model Every consultant gets an assistant. Individual output rises. The delivery model and the pricing stay the same. Margin stays flat.	Rebuild at least one engagement type around the new unit Junior plus agents plus extracted procedures delivers senior-tier output. One proof point changes the internal conversation.
Wait for the bench to develop through existing pathways Junior consultants without routine work have nothing to build judgment on. The 2030 senior bench thins before anyone names it.	Redesign the junior entry around judgment from day one New consultants work on decisions and exceptions, paired with agents. Judgment builds faster than through repetition alone.

Same talent. Same twelve months. Opposite outcomes. The difference is what the delivery engine is built to do with the efficiency, not whether the efficiency exists.

THREE QUESTIONS

Delivery Margin. Technical Credibility. Bench Velocity.

Three questions every IT consulting COO is tracking. None of them used to be the same question. They are now.

01 · DELIVERY MARGIN

Why is our margin drifting on work our rate card says should hold?

The rate card has not changed. Utilisation is within target. But realisation on mid-level engagements is sliding, and client procurement is arriving with AI-efficiency benchmarks on the RFPs where the comparison is clearest. The tools your clients already own are delivering the analysis work that justified the junior rate. The rate card is fine. The pricing model is the question.

02 · TECHNICAL CREDIBILITY

Does our delivery approach still mark us as the technology experts?

Your firm is supposed to be ahead on technology. If your delivery model in 2026 looks structurally similar to 2022, clients in procurement and technical leadership notice before they say anything. The firm that is visibly ahead on AI-augmented delivery is not just more efficient. It is more credible in the room where the next engagement is being scoped.

03 · BENCH VELOCITY

What happens to our senior bench if the junior training ground disappears?

The routine work that turned your junior cohorts into reliable mid-level consultants in three to four years is now running on agents. No routine work, no training volume. The bench thins two to three years after the routine work disappears. Most delivery COOs are not tracking bench velocity as a metric yet. The ones who are have already started rebuilding the entry pathway.

What the firms winning this play are actually doing

Four concrete moves any IT consulting COO can start this quarter. Each touches one part of the delivery engine. Together, they become one architecture the firm runs on.

1 **Delivery: decompose every engagement type to the deliverable task, then price around the tasks that still make money**

Most COOs are redesigning engagement types. The right unit is the deliverable task inside the engagement. List the task bundles across your standard engagements. Mark each: agent-owned with human audit, composite delivery, senior-only, or retired. What remains is a different engagement at a different cost base. The COO who walks into the next pricing conversation with the task-level map owns it. The ones bringing engagement-category P&Ls are running the 2020 model.

2 **Knowledge continuity: extract senior consultant knowledge as a side-effect of active engagements, not as a documentation project**

Knowledge-management initiatives have failed in every consulting firm for twenty years. The judgment is not easily articulable in prose. Extract from the artefacts seniors already produce: architecture decision records, project retrospectives, client-specific notes, rejected-alternative logs. The firm captures reasoning patterns from paid work. The senior is not asked to explain themselves in a separate exercise. The knowledge stops being biographical.

3 **Bench velocity: redesign the junior entry pathway around judgment work from day one, not routine work**

New consultants do not start on boilerplate. They start on exception-handling, design decisions, and client-facing work, paired with agents that carry the substrate. The four-year judgment curve compresses to two. The junior who in 2022 built competence through repetitive coding tasks builds it in 2026 through high-volume decision work. A different kind of competence, formed faster, much harder to replicate elsewhere.

4 **Technical credibility: build AI-augmented delivery as a visible capability the client sees inside the engagement**

The credibility argument is won in the delivery, not in marketing. The client who watches your team work differently, who sees an agent-augmented architecture review produce outputs faster and with clearer rationale, has a credibility data point no case study can replicate. One engagement redesigned around the new unit, delivered transparently, does more for your market position than any capability presentation.

The only firm that joins delivery-floor judgment to working infrastructure

The IT consulting COO has four places to go for help. None of them can do what we do.

The big consulting firms' operations practices will sell a €400K to €1.5M diagnostic delivered by partners billing daily rates. They produce operating-model decks. They cannot put working delivery infrastructure into your engagement workflow in three months. Their own revenue model is built on the same hourly billing your problem is about.

The HR and learning development platforms will sell a talent-development programme with certifications. They develop individual consultants. They do not redesign delivery economics or the pricing model that determines whether that development translates into margin.

The hyperscaler consulting arms will build AI tools tuned to their own platform. Efficiency your consultants gain becomes a dependency on their infrastructure, not an asset your firm owns.

Your internal IT function and workflow automation vendors automate specific steps in your delivery pipeline. They sit too far from how your seniors actually reason to capture the judgment that makes your delivery distinctive. The automation runs. The knowledge problem persists.

The market offers four things you have already said no to. GRAIL is the fifth option.

WHAT YOU GET

Senior advisory that speaks delivery-floor, not strategy-deck

Working infrastructure your consultants use the same week

Custom tools built on your engagements, your clients, your domain

The start of a knowledge infrastructure your best seniors feed

No vendor dependency at any phase of the engagement

WHY IT MATTERS

Founded by practitioners who have run consulting delivery

Business model first, technology second

Three weeks from first meeting to working infrastructure

Your team keeps the infrastructure; you own the IP

Every next step earned by what just shipped, never assumed

Three sessions with your delivery team. Working tools from week one.

We start with your delivery leadership because the change cannot outrun the team's personal understanding of what is possible. Three workshops. Each delivers standalone value. Each earns the right to the next.

WORKSHOP 01	WORKSHOP 02	WORKSHOP 03
How to think	Your delivery infrastructure	Ready to run
Walk through what is happening in IT consulting as a sector and inside your delivery engine. Build the first piece of working infrastructure in the room: a client-specific agent for one of your real engagements, extracted from the project archive you already have. Your delivery team leaves with a shared view and the first piece of working infrastructure already in their hands.	A half-day build. Your Head of Delivery, Head of Practice, Head of Architecture, and Head of Talent each leave with five to ten pieces of working infrastructure. Engagement-scoping agents. Client-knowledge capture tools. Architecture decision records. Delivery quality review agents. Tools your teams use the same week, not decks about tools.	Refined infrastructure from real use. The rollout plan for your wider delivery organisation. The knowledge-extraction plan for your senior consultants. The junior pathway redesign. The CEO and CFO conversation with numbers they can check: delivery margin by engagement type, realisation by pricing model, senior-relationship risk by account.

Three weeks. No IT project. No board approval needed to start. The infrastructure your delivery engine is already running on when we finish.

FIVE QUESTIONS

For your next delivery review

The questions where your delivery team cannot agree on an answer are the ones worth an hour on the agenda. Ask these honestly.

Of the last ten renewal conversations where delivery margin came in under plan, how many involved a client arriving with AI-efficiency benchmarks your team had not anticipated? If you do not know, how would you find out?

If your five most senior consultants each left for client-side roles next quarter, what would happen to the accounts they carry? What have you actually transferred to the rest of the bench in the last twelve months?

Can you point to one engagement type in your delivery model that is measurably faster, at higher quality, and at better margin this quarter than six months ago? If not, what are the next four quarters going to add up to?

What percentage of your junior consultants' time is currently spent on work that builds judgment rather than work that could be agent-owned? And what would it take to change that ratio this quarter?

Which of your mid-level consultants is already working the way the rest of the bench needs to work by 2028? What would it take for that to be the default across your delivery organisation?

If any of these produces a disagreement in your delivery team, that is the conversation worth having before the board asks.

JOHAN GRÖNSTEDT · GRAIL · JOHAN@GRAIL.WORKS · GRAIL.WORKS

IN THIS SERIES

There's a version of this briefing for every seat at the table.

GRAIL has produced industry briefings calibrated to the specific reality of each C-suite role.

CEO

Revenue. Margin. Valuation.

Three numbers your board reads. AI is rewriting all three in your industry.

CFO

Capital. Forecast. Story.

What you allocate this year is the story your board reads next year.

COO

▶ YOU ARE READING

Cheaper. Better. Faster.

The triple threat that used to be a tradeoff is now table stakes.

CHRO

Capacity. Capability. Promise.

The biggest talent transition of a career. Compete on capacity, not headcount.

CPTO

Moat. Speed. Allocation.

The platform moat that survives 2028 is being chosen this year.

CCO

Margin. Pace. Position.

The whole GTM playbook is being rewritten in real time, across every commercial team at once.

Visit grail.works for the full series.

GRAIL

An AI-native management consultancy helping European mid-market leadership teams figure out what fundamentally changes about their business as AI enters their industry.

www.grail.works