

Capability OS

Not a course catalogue. A behavioural architecture: four domains, 12 clusters, measured identically from diagnostic to review.

ABOUT AZIMUT

We don't sell workshop attendance. We build capability that holds under pressure.

Since 2008 we have worked with teams and managers across Romania and the wider region. We learned that a good exercise opens a conversation, but only a system changes behavior: a diagnostic, facilitated labs, practice between sessions, and reinforcement from the direct manager. The activities in this guide are the front door — the easy, useful part. The rest is architecture.

THE NUMBERS BEHIND THE WORK

20,000+

participants trained

9.27/10

average trainer score

600+

programs delivered

18

years of applied practice

DELIVERED FOR TEAMS AT

Bosch · Continental · Banca Transilvania · Endava · Deloitte · Bombardier · ZF · Mazars · Groupama · Betfair · Steelcase · Gameloft

WHERE THIS STARTS

Training gets bought as content and delivered as an event

You probably have the feedback forms and the good scores already. People leaving the room energised. And three months later, the same meetings where nothing is decided, the same missed commitments with no consequence, the same disagreements aired in the corridor rather than in the room.

We have seen this often enough to stop blaming the provider or the participants. A working year has roughly 250 working days. A conventional programme occupies two of them.

248

The days where behaviour actually gets shaped: Monday's meeting, the difficult conversation postponed for three weeks, how a manager reacts when someone tells them the deadline is not realistic.

Awareness is created in the room. Behaviour is formed across the other 248 days. A programme can score extremely well on satisfaction and change nothing about how the team works. The two are not in conflict — they measure different things: one measures the experience inside the room, the other should measure behaviour outside it.

WHAT THE RESEARCH SAYS

- ▶ Ebbinghaus documented in the nineteenth century how quickly unpractised information disappears. The forgetting curve is not a marketing metaphor, it is a design constraint: without spaced repetition, content evaporates no matter how good it was.
- ▶ Kirkpatrick, and later Brinkerhoff, showed that most training evaluation stops at reaction ("they liked it") and learning ("they understood it"), and never reaches behaviour ("they do it") or result. What is not measured does not get designed for.
- ▶ Gallup has consistently found that the direct manager explains roughly 70% of the variance in team engagement. Any intervention that routes around the direct manager is working against the strongest factor in the system.

THE DIAGNOSIS

This is an architecture problem, not a content problem and not a motivation problem

When a programme fails to produce change, the first three explanations we hear are almost always the same three. In our experience, all three point in the wrong direction.

THE EXPLANATIONS THAT DO NOT HOLD

- ▶ **"The content was not right."** Leadership content is, for the most part, a public good. Feedback models, delegation frameworks, listening techniques — they are published, tested, freely available. If information were the limiting factor, the problem would have solved itself years ago.
- ▶ **"People were not motivated enough."** Motivation explains the start, not the consistency. Nearly every participant leaves the room genuinely intending to work differently. That intention survives until the first overloaded week.
- ▶ **"They were not the right people."** The same people, placed in a system with explicit expectations and weekly reinforcement, behave differently. When behaviour changes with the context, the cause was never in the people.

What is missing, almost every time, is the **transfer layer**: the mechanism that carries a behaviour out of the room and into the work, and holds it there until it becomes default.

The content is not the problem. The architecture around the content is the problem. An event gets purchased and a system gets expected.

THE CAPABILITY OS THESIS

WHAT THE TRANSFER LAYER ACTUALLY CONSISTS OF

- ▶ A stable taxonomy of behaviours, identical from diagnostic through to reporting.
- ▶ Distributed practice between sessions, calibrated to each person's role.
- ▶ A direct manager who observes, names and corrects that behaviour, week after week.
- ▶ Measurement of behaviour rather than satisfaction, against a baseline taken before the programme starts.

Capability OS is our answer to that problem. It does not add more content. It adds the infrastructure that gets content out of the room and into the work.

01

THE PRODUCT

Capability OS

Four interdependent domains. Twelve clusters of observable behaviour. One vocabulary, from the diagnostic through to the final report.

WHAT IT IS

A standardized behavioural architecture, not a course catalogue

Until last year we sold a catalogue of 25 programmes. We retired it. Not because the programmes were weak — the scores said otherwise — but because catalogue logic prevents the one thing a development partner is actually paid for: accumulation.

In a catalogue, every programme arrives with its own model, its own language and its own measurement. "Assertive communication" in spring does not connect to "Effective feedback" in autumn, and neither connects to the engagement survey the L&D (Learning & Development) team runs in parallel. What you end up with is a file of delivered programmes and no trajectory.

Capability OS replaces the catalogue with a frame of reference: **four domains** — Clarity, Accountability, Dialogue, Learning — split into **twelve clusters** of observable behaviour. You do not pick topics from a list. You start from a diagnostic that shows where the system gives way, and you intervene on those clusters.

What the difference looks like in practice

WHAT IT IS

- One vocabulary, used from the diagnostic through to the report to leadership.
- Behaviour described concretely enough that two different managers recognise the same thing.
- Practice spread across weeks, not content compressed into two days.
- The direct manager involved weekly, with one small, clear task.
- A baseline measurement, agreed to before anyone knows how it will look.

WHAT IT IS NOT

- Topics picked from a catalogue against whatever budget is left in December.
- A new model from every provider and in every financial year.
- A high-energy day followed by a Monday exactly like the last one.
- An attendance report read as an outcome.
- Individual data used to make decisions about people.

TAXONOMY DISCIPLINE

Why the names never change

The same twelve names appear in the diagnostic instrument, in the lab agendas, in the assessment rubrics, in the manager scripts and in the report to leadership. It is a constraint we impose on ourselves and it is the most valuable part of the product.

- ▶ The diagnostic measures exactly the clusters the labs train — not adjacent variables.
- ▶ The manager observes the behaviour described in the same rubric the participant sees.
- ▶ Leadership reads one vocabulary rather than four competing models.
- ▶ Progress becomes comparable across teams, across programmes and across years. A measurement from 2026 sits next to one from 2029.

Domain and cluster names stay in English in every delivery language. This is not affectation: a single untranslated vocabulary is the condition for data staying comparable between teams working in different languages, and from one year to the next. Content, exercises and cases change with context. The structure does not.

THE TAXONOMY

Four domains, twelve clusters

DOMAIN	THE THREE CLUSTERS	WHAT GIVES WAY WHEN THE DOMAIN IS WEAK
Clarity	Priority Translation Decision Ownership Expectation Framing	People work hard in different directions. The same decisions get reopened monthly because nobody knows who holds them. Expectations are set in adjectives and then assessed in numbers.
Accountability	Commitment Discipline Ownership & Follow-Through Performance Correction	Commitments become optional. Deadlines slip with nobody remarking on it. The correction conversation is postponed until it becomes an exit conversation.
Dialogue	Active Listening & Understanding Constructive Dissent & Conflict Navigation Feedback Exchange	Disagreement moves to the corridor. Decisions look unanimous in the meeting and get relitigated privately. Feedback travels downward only, once a year.
Learning	Reflection & Sensemaking Experimentation & Adaptive Action Knowledge Integration & Team Learning	The organisation repeats the same mistakes with new people. What one person learned stays in their head. Retrospectives turn into status reporting.

The four domains are not independent modules and we do not sell them separately as "areas of interest". They are interdependent, and imbalance between them produces predictable symptoms: Accountability without Dialogue becomes control; Clarity without Learning becomes rigidity; Dialogue without Accountability becomes a pleasant conversation with no consequence. This is why the diagnostic measures all four domains even when the intervention starts in one.

What Clarity and Accountability actually train

Clarity

A team's capacity to know what matters now, who decides, and what "done well" means. Without Clarity, the other three domains work efficiently towards the wrong objective.

- ▶ **Priority Translation** — turning strategy into two or three priorities the team can repeat from memory, without opening the deck.
- ▶ **Decision Ownership** — who decides what, with what mandate, by when; removing the decisions that belong to nobody.
- ▶ **Expectation Framing** — stating expectations as observable behaviour and checkable standards rather than adjectives.

Accountability

Not "holding people to account", but making a commitment a unit of work you can rely on. This is the domain most often mistaken for toughness, and most often treated with reporting systems that change nothing.

- ▶ **Commitment Discipline** — explicit commitments with an owner and a date, tracked as routine and without drama.
- ▶ **Ownership & Follow-Through** — carrying work to completion including when nobody is watching any more.
- ▶ **Performance Correction** — the correction conversation held early, specifically, without humiliation, and ending in a verifiable agreement.

THE COMMON CONFUSION

Accountability is not solved with a dashboard. A report shows who was late; it does not produce the conversation that changes what happens next time. When an organisation adds a third reporting tool for the same problem, the missing thing is the conversation, not the visibility.

What Dialogue and Learning actually train

Dialogue

This is where it gets decided whether bad news reaches the right person in time. Google's Project Aristotle identified psychological safety as the strongest predictor of team effectiveness, and Amy Edmondson's work makes clear that psychological safety is not comfort.

Psychological safety is not the comfort of agreeing. It is the willingness to say the uncomfortable thing in the room, rather than in the corridor afterwards.

THE DIALOGUE DOMAIN

- ▶ **Active Listening & Understanding** — understanding before responding; explicitly checking comprehension, especially under time pressure.
- ▶ **Constructive Dissent & Conflict Navigation** — disagreement raised in the meeting rather than the corridor; conflict conducted on substance rather than on people.
- ▶ **Feedback Exchange** — feedback in both directions, as a weekly practice rather than an annual event.

Learning

The domain that decides whether an organisation accumulates capability or merely accumulates tenure. It is also the easiest to sacrifice under pressure, because its absence produces no immediate crisis — it produces the same crisis again every two years.

- ▶ **Reflection & Sensemaking** — stopping deliberately to make sense of what happened, before moving to the next agenda item.
- ▶ **Experimentation & Adaptive Action** — small experiments with a hypothesis, a success criterion and a stop criterion.
- ▶ **Knowledge Integration & Team Learning** — what one person learned becoming team practice rather than an anecdote over coffee.

02

DELIVERY

The delivery spine

Five layers. The first three produce capability, the fourth keeps it alive, the fifth makes it visible.

THE DELIVERY SPINE

Five layers, in the order that matters

The order is not a calendar preference. Each layer assumes the one before it was done: without the diagnostic we train what seems plausible, and without reinforcement we train for nothing.

01

Diagnostic

A ranking of the clusters constraining the system now.

02

Labs

The behaviour practised on real situations, corrected in the moment.

03

Practice

Short repetition between sessions, calibrated to the role.

04

Reinforcement

The behaviour observed and named by the direct manager, weekly.

05

Review

Measurement against the baseline, and the decision for the next cycle.

- 1 **Diagnostic.** We measure the starting point across the twelve clusters: scenario-based assessment, the manager's and the team's perspective, plus organisational context. The output is not a sixty-page report but a ranking — which two or three clusters are constraining the system right now.
- 2 **Facilitated labs.** Working sessions, not lectures. Participants practise the cluster behaviour on situations from their own organisation, with facilitation, correction in the moment, and an explicit agreement about what they will attempt before the next session. The facilitator's job is to provoke real practice, including the conversations people avoid.
- 3 **AI-personalized practice between sessions.** Between labs, each participant receives short scenarios calibrated to their role, their starting level and the cluster in play, at a cadence that works against the forgetting curve. Practice takes minutes, not hours, and fits the rhythm of real work.
- 4 **Manager reinforcement — roughly 15–20 minutes per week.** The direct manager observes one cluster behaviour, names it explicitly, and either corrects or confirms it. They do not run training. They get a prepared question, a one-page rubric, and somewhere to record what they saw.
- 5 **Measurement and review.** The same scenario assessment, the same rubrics, compared against the baseline, alongside manager-observed behaviour and the business signals the client chose. The findings decide what gets trained in the next cycle.

LAYER FOUR

The direct manager is the load-bearing layer

When budget or calendar tighten, the first layer to disappear is the fourth. It is also the one without which the rest quietly stops meaning anything.

70%

Of the variance in a team's engagement is explained by the direct manager, in Gallup's data. A programme that routes around them is working against its own outcome.

A programme without manager reinforcement is not a cheaper programme. It is a different programme, one that produces awareness instead of behaviour. We would rather deliver a single cluster with the reinforcement layer than four clusters without it, and we say so before the contract.

WHAT THE MANAGER ACTUALLY DOES**The 15–20 minute ritual**

- ▶ One prepared question, added to a one-to-one that already exists — not a new meeting.
- ▶ Observing a single behaviour from the cluster in play, in real context.
- ▶ Naming it explicitly: what they saw, what effect it had, what they would change next time.
- ▶ A short note in the rubric, which feeds the measurement at the end of the cycle.

WHAT WE DO NOT ASK OF THE MANAGER

- ▶ To run training. We ask them to observe and to name, not to teach — that distinction is what makes the ritual survivable.
- ▶ To appraise people from their notes. The rubric measures behaviour inside the programme, not performance against a job description.
- ▶ To invent exercises. The question and the rubric arrive prepared, for the cluster in play.
- ▶ To find extra time. The ritual sits inside a conversation that already happens every week.

THE ROLE OF TECHNOLOGY

Where AI fits, and where it stops

AI personalizes practice between sessions. That is the whole remit. It does not facilitate, it does not assess people, and it does not make decisions that affect anyone's career.

The between-sessions practice layer is the one that simply does not exist in a conventional programme, because it cannot be scaled by hand. We cannot put a facilitator next to every participant every week. This is where the technology solves a genuine architecture problem rather than a marketing one.

WHAT IT DOES

- ▶ Generates practice scenarios from the participant's role, their level on the cluster, and the organisational context.
- ▶ Adjusts difficulty based on how they respond, so practice stays at the edge of competence rather than below it.
- ▶ Holds the cadence between sessions — spaced repetition, which is the direct countermeasure to the forgetting curve.
- ▶ Prepares the manager's conversation: what the team practised this week and what is worth observing.
- ▶ Aggregates anonymised patterns for the facilitator, so the next lab starts from where people got stuck rather than from the planned agenda.

WHAT IT DOES NOT DO, BY DESIGN

- ▶ **It makes no personnel decisions.** No output of the system feeds promotion, bonus or exit decisions.
- ▶ **It does not replace the facilitator.** Conflict, dissent and performance correction are learned in a room, with someone who can read what is not being said.
- ▶ **It does not hand individual scores to leadership.** Individual practice data stays with the participant; the organisation sees it aggregated by team and by cluster.
- ▶ **It does not invent content.** Scenarios are generated inside the validated cluster rubrics, not freely.

THE QUESTION TO ASK ANY PROVIDER

If someone tells you their AI replaces the facilitator, ask who runs the conversation in which two managers genuinely disagree, with something at stake, in the same room. The answer tells you everything you need to know about the product.

MEASUREMENT

Behaviour, not satisfaction

"I liked the training" is not an outcome. It is a hygiene condition: if it is missing we have a problem, and if it is present we have proved nothing. We measure four things, and none of them is the mood in the room.

- 1 **Scenario-based assessment, read against cluster rubrics.** Participants do not answer questions about theory. They respond to realistic situations — a missed commitment, a disagreement in a meeting, a priority landing on top of three others — and the response is read against that cluster's rubric, with levels described as behaviour rather than as grades.
- 2 **A baseline, before and after.** The same assessment, the same rubrics, at the start and at the end of the cycle. Without the "before" measurement, any number at the end is an assertion rather than evidence.
- 3 **Manager-observed behaviour.** The weekly notes from the manager ritual aggregate: did the behaviour appear, in what context, at what frequency. It is the only source that measures what happens across the 248 days.
- 4 **Business signals the client chooses.** We do not propose these. The client picks two or three indicators they already track — retention in critical roles, time from decision to execution, escalation rate, absence, the internal engagement survey — and we read them alongside the behavioural data.

WHAT YOU WILL NOT HEAR FROM US

We do not claim causality on business indicators. Too much moves in parallel inside an organisation for anyone to separate the effect of one programme honestly. We track them as context signals, and we make claims about effect only at the level where we genuinely measure: behaviour.

WHAT WE DO NOT CALL AN OUTCOME

- ▶ The satisfaction score at the end of the session.
- ▶ Attendance or completion rate.
- ▶ Hours delivered.
- ▶ A stated intention to apply something.

FORMATS

Four formats, and how to choose between them

All four use the same taxonomy and the same rubrics. They differ in how much of the architecture makes it into delivery — and therefore in how much of the behaviour survives afterwards.

FORMAT	DURATION AND COVERAGE	WHO IT SUITS	BEST WHEN
Lab	One facilitated session, half a day to a day. One cluster.	One team and its direct manager.	The problem is clearly bounded, the manager will run the weekly ritual afterwards, and you want to see the method working before a larger commitment.
Sprint	4–6 weeks. Three clusters — one complete domain.	A management layer with the same weakness.	The symptoms gather inside a single domain — Accountability, for instance, after a reorganisation left commitments without owners.
Accelerator	8–12 weeks. Two domains, six clusters.	Several teams inside the same structure.	Two domains are blocking each other — most often Clarity and Accountability — and you need full pre/post measurement rather than an impression.
Pilot	3–6 months. Full Capability OS, one pilot group.	Organisations building an internal development standard.	The real question is what your leadership standard looks like in three years, and the architecture has to be validated on one group before it is extended. Includes the full diagnostic, manager preparation for the reinforcement layer, and reporting to leadership.

HOW WE CHOOSE TOGETHER

- ▶ We start from the diagnostic, not the format. The format is a consequence of the cluster ranking, not a calendar preference.
- ▶ If the problem is one behaviour, in one team, with an engaged manager — **Lab**.
- ▶ If the symptoms group inside a single domain and recur across several managers — **Sprint**.
- ▶ If two domains are blocking each other — **Accelerator**.
- ▶ If the real question is "what does our leadership standard look like in three years" — **Pilot**.
- ▶ A standalone Lab has no between-sessions practice layer. We recommend it as an entry point or as a module inside a larger programme, not as a fix for a systemic problem. We say that before the contract, not after.

SCOPE AND BUDGET

Why there are no prices in this brochure

Not out of coyness. Because a figure lifted out of context tells you less than it appears to, and this document should still be accurate in two years.

The same format can mean very different projects. What moves the number is how many groups take part, which language delivery runs in, how wide the diagnostic goes, how many direct managers we prepare for the reinforcement layer, and over how long the whole thing runs. So scope and budget get set in a scoping conversation at the start, not in a price list.

If the budget is fixed, this is the order of priorities



The proportions are a starting point, not a rule — they shift with how many teams take part and how far the starting point is from where you want to be. What we do not recommend is cutting the middle slice to buy more days in the room. That gives you a bigger programme that holds for less time.

What it helps us to know in advance

- ☐ How many teams and how many direct managers would be in scope.
- ☐ Who owns the outcome internally, and who sponsors it at leadership level.
- ☐ Which language delivery runs in, and whether the groups are mixed.
- ☐ What has already been tried on the same problem, and what did not hold.
- ☐ What else sits in the organisation's calendar during the months the programme would run.
- ☐ The two or three indicators you already track and would read alongside the behavioural data.

We leave that conversation with three things: what we would measure, what we would train first, and in which format — plus a budget range put on the table there, rather than after three rounds of email.

FIT

Who this works for, and who it does not

A capability programme asks something of the organisation too, not only of the provider. We would rather test that in the first conversation than in month three.

IT WORKS WHEN

- ▶ Direct managers can commit 15–20 minutes a week, and their own bosses treat that as work rather than as an extra.
- ▶ Someone inside the organisation owns the outcome — usually in HR or L&D — with a leadership sponsor who will field uncomfortable questions.
- ▶ There is willingness to measure the starting point, including the risk that the first measurement looks bad.
- ▶ The horizon is months, not days. Behaviour does not recalibrate over a weekend.
- ▶ It is accepted that individual data stays with the participant and the organisation reads aggregates.

WE ARE NOT THE RIGHT PARTNER IF

- ▶ You want a one-day workshop that will "fix communication". We can deliver an excellent day; we cannot promise it changes anything on Monday, and we will not pretend otherwise.
- ▶ You are looking for a motivational speaker for the annual event. That is a legitimate product, just not ours.
- ▶ You want to license content and materials without facilitation. A taxonomy without the transfer layer is precisely the problem we are trying to solve.
- ▶ Direct managers cannot or will not take part. Without them, the heaviest factor in the system is absent.
- ▶ The expected outcome of the programme is a satisfaction score or a number of reported hours.
- ▶ You want individual participant data to inform personnel decisions.

We say no a few times a year. Every time, the reason is one of the above — and every time, the conversation turned out more useful than a badly matched contract.

WHAT SITS BEHIND THE METHOD

Since 2008 600+ programs 20,000+ participants 9.27/10 average trainer score

Capability OS did not appear on a whiteboard. It is the shape our work took with teams at organisations such as Bosch, Continental, Banca Transilvania, Endava, Deloitte, Bombardier, ZF, Mazars, Groupama, Betfair, Steelcase and Gameloft — after seeing, often enough to be sure, which programmes still hold at six months and which do not.

NEXT STEP

We start with the diagnostic, not the proposal

We cannot recommend a format before we know which clusters are constraining the system. The first conversation takes about 45 minutes and ends with one of three conclusions: what we would measure, what we would train first, or why we are not the right partner.

- ▶ Book a diagnostic conversation — azimut.team/en/contact
- ▶ See the full taxonomy and the four formats — azimut.team/en/capability-os
- ▶ Start with the diagnostic instrument — azimut.team/en/capability-snapshot

Azimut · Capability Development Partner ·

NEXT STEP

Want to see what this looks like for your organization?

azimut.team/en/capability-os →

We publish one capability insight a month — a single useful idea about how behaviour actually changes in organisations. Every guide, free, plus the subscription:

azimut.team/en/insights →



NEXT STEP

Tell us the behaviour you need to change.

A 45-minute diagnostic, no pitch deck. You'll leave with a sharper read on the gap — whether or not you work with us.

azimut.team/en/contact →