



COMPANION GUIDE

The Cognitive Delivery Index

A self-assessment instrument for governed AI participation in software delivery

Jason Cruz Falzon
Creator of the Cognitive Delivery Framework

Version 1.0
cognitivedelivery.co.uk



What this guide is for

Most teams adopting AI agents in software delivery can already point to a governance document. Almost none of them can demonstrate that the document is doing any work. The Cognitive Delivery Index exists to surface that gap before it becomes an incident.

The Cognitive Delivery Framework starts from a simple observation. The software development lifecycle was designed for a world where the humans doing the work were the same humans being held accountable for it. Agents have broken that assumption, and most teams have responded by writing policies that describe how things should work rather than measuring how things actually work.

The Index is a measurement instrument, not a maturity model. It does not tell you whether you are good or bad. It tells you, on six specific dimensions, whether the controls you have on paper are doing anything in practice. The most important finding it tends to produce is not a low score. It is the discovery that a team scoring four out of five (Complete) is often only one observed behaviour away from scoring two (Informal), because the documentation has never been tested under pressure.

That gap, between what is documented and what is actually exercised, is the central concern of this guide. I call it governance theatre, and the Index is built to find it.

"A control that has never been tested is a hypothesis, not a control."

This document is the long-form companion to the one-page CDICanvas. The Canvas is for the assessment session. This guide is for what comes before and after: understanding the dimensions, calibrating your scoring, facilitating an honest conversation with your team, and turning the assessment into ninety days of focused work.

HOW TO USE THIS GUIDE

Read sections one to four before your first assessment. Use sections five to ten as the dimension reference during scoring. Return to sections eleven onwards once you have a completed Canvas in front of you and a decision to make about where to invest.

What is in this guide

01	What the CDI measures, and what it deliberately does not	04
02	The six dimensions, in plain terms	05
03	The five levels and the governance theatre gap	06
04	Dimension one: Provenance Integrity	07
05	Dimension two: Supervisory Clarity	09
06	Dimension three: Scope Definition	11
07	Dimension four: Review Substance	13
08	Dimension five: Constraint Fidelity	15
09	Dimension six: Capability Calibration	17
10	Running the assessment session	19
11	Reading your results	20
12	From canvas to ninety-day action	21
13	Common failure modes	22
14	Re-assessment cadence	23
15	About the framework	24

What the CDI measures, and what it does not

The Index measures the quality of governance over agent participation in software delivery. It does not measure agent productivity, code quality, or whether the team is using AI well. Those things matter. They are not the same thing.

What the CDI measures

The CDI measures six dimensions of governance, each scored from one to five, against a single question. Can you demonstrate that the control is real?

A control is real when it produces observable behaviour under pressure. A documented escalation path that has never been used in eighteen months of agent work is not a real control. A scope boundary that lives in a Confluence page but is not enforced by the agent's runtime is not a real control. The Index draws a sharp line between description and demonstration, and that line is where most teams discover their actual position.

What the CDI does not measure

The Index is silent on whether your agents are good, fast, or appropriate for the work. A team running excellent agents with reckless governance will score badly. A team running mediocre agents under tight governance will score well. This is intentional. Other instruments exist for measuring agent capability. The CDI exists because the agent capability conversation has consistently outpaced the governance conversation, and the Framework's position is that capability without governance is a liability dressed up as productivity.

Who should run an assessment

The Index works at the level of a delivery team, a programme, or a portfolio. It is not designed for the level of an individual contributor or a single project. Below the team level, the dimensions collapse into noise. Above the portfolio level, the conversation becomes too abstract to act on.

The right people in the room are those who would be asked to explain an agent-related incident to a regulator or a board. If those people cannot answer the six diagnostic questions with evidence, you have a CDI problem, regardless of the score on the day.

A NOTE ON HONESTY

The Index only works if the people scoring it are willing to mark themselves down. Optimistic scoring produces a number that flatters the team and changes nothing. The most useful sessions are the ones where someone says "I know we have that document, but I have never seen anyone actually do that," and the room agrees.

The six dimensions, in plain terms

Each dimension answers a single diagnostic question that a senior leader should be able to answer with evidence at any time. If they cannot, the dimension is not yet at level four, regardless of what your governance documents claim.

ONE Provenance Integrity	Can you reconstruct what agents produced and why? If a defect, a regulator, or a curious engineer asked which lines of code were generated by an agent, with what prompt and what supervisor approval, could you answer? Provenance is the foundation of every other control, because without it, accountability is theatre.
TWO Supervisory Clarity	Who is accountable for agent output, with what authority? When agent work goes into production, a named human carries responsibility. That human must have the authority to refuse, the time to engage, and the standing to escalate. If supervision is implicit or distributed across "the team", supervision is not happening.
THREE Scope Definition	Do agents operate within explicit boundaries? An agent permitted to do anything will eventually do something it should not have. Scope must be defined, enforced technically where possible, and reviewed when the agent's behaviour suggests the boundary is in the wrong place.
FOUR Review Substance	Is review of agent work substantive or procedural? A reviewer who clicks approve on every pull request is not reviewing. The CDI measures whether reviewers reject things, push back on things, and have the time and incentive to engage with what they are signing off.
FIVE Constraint Fidelity	Are rules explicit enough for agents to follow? Agents cannot infer constraints that humans hold tacitly. If your team's coding standards live in someone's head, the agent will not respect them. Constraint fidelity is the discipline of making the implicit explicit and machine-readable.
SIX Capability Calibration	Does expansion follow evidence or enthusiasm? Most teams expand agent scope because something worked once or because a stakeholder is excited. Calibration asks whether expansion decisions rest on observed performance, defect rates, and supervisor input, or on optimism.

The dimensions are intentionally orthogonal. A team can score five on one and one on another, and that pattern is itself diagnostic. A team that is excellent at provenance but weak at review substance is collecting evidence it never uses, which is a specific and recognisable failure mode worth addressing on its own terms.

The five levels and the governance theatre gap

Each dimension is scored from one to five. The scale is not linear. The hardest jump on the scale is from four to five, and that jump is where the Index does its most useful work.

Level 1	Level 2	Level 3	Level 4	Level 5
Absent	Informal	Documented	Complete	Substantive
Nothing exists. The control is not present in any form. The team has not yet considered the question.	The control exists by convention. People know roughly what is supposed to happen, but it lives in conversations and assumptions.	The control is written down. Coverage is partial and consistency is uneven, but a document exists that can be pointed at.	The control is fully documented and procedurally adopted. On paper, everything is in place. Auditors are satisfied.	The control produces observable behaviour. It is exercised, tested, and visibly shapes decisions. The control is real.

The governance theatre gap

The gap between level four and level five is the central insight of the Index. Most governance maturity models stop at four. They define maturity as the condition of having all the documents, all the policies, all the role descriptions. The CDF position is that this is the condition in which most governance failures occur, because the existence of a document is consistently mistaken for the existence of a control.

A team at level four can pass an audit. A team at level five can demonstrate, with examples from the last quarter, that the control changed an outcome. Those are not the same thing. The first is sufficient for compliance. Only the second is sufficient for safety.

WHAT GOVERNANCE THEATRE LOOKS LIKE

An escalation policy that has never been escalated. A review checklist that is signed but not engaged with. A scope document that lives outside the agent's runtime context. A capability committee that approves every request put to it. These are the signs of a level-four organisation that believes itself to be at level five.

Why the jump from four to five is the hardest

Reaching level four requires writing things down. Reaching level five requires changing how people behave, where their time goes, and which decisions get challenged. The first is a documentation exercise. The second is a cultural one, and culture is harder to move than process.

This is why a team's level four to five distribution often matters more than the headline score. A team at four across the board is in a more dangerous position than a team scoring three on most dimensions and five on one or two, because the first team has bought confidence on credit.

Provenance Integrity

Can you reconstruct what agents produced and why?

Why this matters

Provenance is the precondition for every other dimension. Without a reliable record of what an agent produced, what was asked of it, and who approved the result, accountability collapses into guesswork. When an incident occurs, the first question is always "how did this code get here". If the answer is "we are not sure", the team has a Provenance Integrity problem before it has any other kind of problem.

This dimension is also the easiest to fake. A team can claim full provenance because every commit has an author, without recognising that author tags do not distinguish human from agent contributions, and that prompts are not retained anywhere. The honest question is not whether you have provenance, but whether you could answer the reconstruction question for last week's deployments in under an hour.

Scoring rubric

LEVEL 1 Absent	No record exists of which contributions were agent-produced. Commits are indistinguishable. Prompts are not retained. No one on the team could reconstruct an agent's reasoning for any specific change.
LEVEL 2 Informal	Some developers tag agent commits informally, perhaps in commit messages. Knowledge of which agent did what lives in chat threads or memory. Reconstruction is possible for recent work, by asking the right person.
LEVEL 3 Documented	A process exists for marking agent contributions. Coverage is inconsistent, and the process has not been retrofitted to historical work. Prompts are sometimes captured, sometimes not.
LEVEL 4 Complete	Every agent contribution is tagged. Prompts are retained. An audit trail exists for the supervisor, the agent, the input and the output. The trail is complete on paper.
LEVEL 5 Substantive	Provenance is queryable and routinely queried. The trail is used in incident review, capability calibration, and supervisor performance conversations. The team can show specific examples from the last quarter where provenance changed a decision.

What signals each end of the scale

SIGNALS YOU ARE AT LEVEL FIVE

- You can answer "show me the prompt that produced this code" within minutes, not days.
- Provenance data has been used in a post-incident review.
- Supervisors reference provenance in their review conversations.
- The provenance system has been queried more than once in the last month.

SIGNALS YOU ARE NOT

- The provenance database exists but no one has logged in for weeks.
- Prompts are stored, but in a format that is not searchable.
- Tagging is done at the end of the sprint, retrospectively, by memory.
- Nobody can name the last time provenance was used to answer a question.

One concrete action to move forward

Pick one defect from the last sprint and ask the team to reconstruct the full provenance of the change that caused it. Track how long it takes and what was missing. The exercise will tell you your true score on this dimension faster than any audit will.

Supervisory Clarity

Who is accountable for agent output, with what authority?

Why this matters

Accountability does not survive being distributed. When "the team" is responsible for agent output, no one is, and the practical consequence is that nothing gets refused, nothing gets escalated, and the agent's work goes into production through the path of least resistance.

Supervisory Clarity asks two questions in sequence. First, is there a named human accountable for each piece of agent output. Second, does that human have the authority and the time to actually exercise the role. A named supervisor who is also delivering their own sprint commitments and reviewing twelve agent pull requests a day is, in practical terms, not supervising. They are signing off.

Scoring rubric

LEVEL 1 Absent	No named owner exists for agent output. Review happens by whoever is available. Authority to refuse is unclear or unexercised. When something goes wrong, the question of who should have caught it has no answer.
LEVEL 2 Informal	Implicit supervisors emerge by convention. Senior engineers tend to review agent work, but no role is defined. Authority is assumed rather than granted.
LEVEL 3 Documented	Supervisor roles are defined per agent or workstream, on paper. The document specifies authority and escalation paths. Adoption is uneven, and supervisors have not been formally trained in the role.
LEVEL 4 Complete	Supervisors are named, trained, and granted explicit veto and escalation rights. Time is allocated for the role. The structure is in place.
LEVEL 5 Substantive	Supervisors actively exercise their authority. Veto is used and respected. Escalations happen. The team can show recent examples where supervisor judgement changed an outcome that the agent and the requester both wanted.

What signals each end of the scale

SIGNALS YOU ARE AT LEVEL FIVE

- A supervisor refused or substantially changed agent output in the last month.
- Supervisors push back without fearing for their delivery commitments.
- Time for supervision is protected in capacity planning.
- Escalation paths have been used recently, and the escalations were taken seriously.

SIGNALS YOU ARE NOT

- The named supervisor cannot remember the last time they refused something.
- Supervisors are also the requesters, with no separation.
- Supervision is squeezed in around delivery, not protected as work.
- Escalation paths exist on paper but lead nowhere recognisable.

One concrete action to move forward

Ask each named supervisor for the last three pieces of agent work they pushed back on, and what changed as a result. If the answer is silence, your level on this dimension is lower than you think, and the action is to make supervisory time visible and refusal psychologically safe.

Scope Definition

Do agents operate within explicit boundaries?

Why this matters

Agents do not respect boundaries that are not enforced. A scope statement in a Confluence page is a polite request. A scope enforced by role-based access control, tool restrictions, or sandbox boundaries is a control. The CDI distinguishes the two ruthlessly, because most teams have the first and believe they have the second.

Scope failures are also the category of failure most likely to produce the kind of incident that ends careers. An agent given broader permissions than its work required, that touches a system it should not have touched, will be remembered. The defence "the policy said it should not have done that" carries no weight when the policy was not enforced.

Scoring rubric

LEVEL 1 Absent	Agents operate with whatever permissions were easiest to grant during setup. Scope is defined by accident, usually broadly. No one has audited what an agent could do if it tried.
LEVEL 2 Informal	Conventions exist about what agents should and should not touch. Conventions are held in conversation rather than enforcement. Violations are corrected when noticed.
LEVEL 3 Documented	A written scope exists per agent or workstream. It lives in a document, not in the agent's runtime. Compliance depends on the agent and its supervisors reading and respecting the document.
LEVEL 4 Complete	Scope is encoded technically. The agent cannot exceed its boundaries because the runtime will not let it. Permissions are scoped to least privilege. Sandbox boundaries are in place.
LEVEL 5 Substantive	Scope is reviewed regularly, tested adversarially, and adjusted on evidence. Violations, when they occur, trigger investigation rather than expansion. Scope changes follow the same evidence standard as code changes.

What signals each end of the scale

SIGNALS YOU ARE AT LEVEL FIVE

- An agent has been blocked from exceeding scope by a technical control in the last quarter.
- Scope is reviewed on a defined cadence, and the review has produced changes.
- Adversarial testing is part of the practice, not an event.
- When an agent needs broader scope, the conversation is about evidence, not convenience.

SIGNALS YOU ARE NOT

- The agent has admin permissions because narrower ones were too much trouble.
- Scope is reviewed at agent onboarding and never revisited.
- Scope violations have been worked around rather than investigated.
- No one can name what the agent's runtime would actually prevent it from doing.

One concrete action to move forward

For each agent in production, write down the three things it should never be allowed to do. Then test whether the runtime would actually stop it. The gap between intention and enforcement is your score on this dimension.

Review Substance

Is review of agent work substantive or procedural?

Why this matters

Review is where governance either becomes real or quietly gives up. A reviewer who approves everything is not a reviewer. A review process that produces a one hundred per cent approval rate on agent work is signalling either that the agent is exceptional, which is rare, or that the review is theatre, which is common.

Review Substance is also the dimension where time pressure does the most damage. Reviewers under sprint pressure approve faster. Approval by a tired reviewer at the end of a sprint is not the same control as approval by an engaged reviewer in the middle of one. The Index measures whether the conditions for substantive review actually exist, not whether the review step is in the workflow.

Scoring rubric

LEVEL 1 Absent	No review of agent work, or review is performative. Sign-off is a click. Reviewers do not engage with the substance of what they are approving.
LEVEL 2 Informal	Reviewers approve based on trust in the agent or the requester. Engagement is shallow. "Looks fine" is a common comment. Rejection is rare.
LEVEL 3 Documented	Review checklists exist. They are followed inconsistently. Some reviewers engage substantively; others sign off. The process tolerates both behaviours.
LEVEL 4 Complete	Reviewers are held to a standard. Sign-offs require demonstrated engagement. The checklist is followed. Time for review is allocated.
LEVEL 5 Substantive	Review catches real defects. Reviewers reject substantively and routinely. Time and incentives support deep engagement. The reviewer's judgement carries weight in the team.

What signals each end of the scale

SIGNALS YOU ARE AT LEVEL FIVE

- The rejection rate on agent work is non-trivial and considered healthy.
- Reviewers comment substantively on more than half of agent submissions.
- Review time is protected in planning and respected in practice.
- Reviewer judgement has changed downstream decisions in the last month.

SIGNALS YOU ARE NOT

- Approval rate is approaching one hundred per cent.
- Reviews complete in seconds, on average.
- Reviewers are also delivery contributors with no time set aside.
- The team would be uncomfortable if reviewers started rejecting more.

One concrete action to move forward

Track the rejection rate on agent submissions for one sprint. Discuss the result openly. If it is too low, the question is not how to make reviewers reject more. The question is what conditions would make substantive engagement possible, and whether the team is willing to change them.

Constraint Fidelity

Are rules explicit enough for agents to follow?

Why this matters

Agents do not pick up tacit knowledge. The unwritten rule that the team avoids a particular pattern, the shared assumption that a certain library is preferred, the convention that nobody documented but everyone follows: these are invisible to the agent. Constraint Fidelity is the discipline of making the implicit explicit, then making the explicit machine-readable, then making the machine-readable available to the agent at the moment it is needed.

This dimension is often weak in teams that are otherwise strong, because the work feels mechanical. Writing down what everyone already knows feels redundant until an agent breaks something by not knowing it. The Index recognises that constraint fidelity is among the least exciting dimensions to invest in and among the highest leverage.

Scoring rubric

LEVEL 1 Absent	Constraints live in tribal knowledge. Agents work blind to them. Violations are caught by humans, sometimes, after the fact.
LEVEL 2 Informal	Some constraints are captured in agent prompts or instructions, ad hoc. Capture depends on the requester remembering. Coverage is partial.
LEVEL 3 Documented	Constraint documents exist and are referenced in agent context. Documents drift from practice. Updates are sporadic.
LEVEL 4 Complete	Constraints are versioned, tested, and included in the agent's runtime context. The agent receives the constraints at the moment it acts on them.
LEVEL 5 Substantive	Constraint violations by agents are detected and treated as defects. Constraints evolve based on observed agent behaviour. The constraint set is a living artefact, not a once-written document.

What signals each end of the scale

SIGNALS YOU ARE AT LEVEL FIVE

- Constraints are versioned in source control alongside the code they govern.
- An agent constraint violation has been logged and addressed in the last month.
- The team has updated a constraint because the agent revealed it was unclear.
- New team members are onboarded by reading the same constraints the agent reads.

SIGNALS YOU ARE NOT

- The constraint document was last updated when the team was half its current size.
- Constraints are pasted into prompts manually, with variation.
- "That is not how we do things here" is still a common comment in agent reviews.
- No one is responsible for keeping the constraint set current.

One concrete action to move forward

Take the next agent rejection that involves "we don't do it that way", and ask whether the rule was in the agent's context. If not, write it down, version it, and add it. Repeat until the agent stops being told things in review that it could have been told in advance.

Capability Calibration

Does expansion follow evidence or enthusiasm?

Why this matters

Most agent scope expansion happens because someone saw a demo. Calibration is the practice of insisting that expansion decisions follow the same standard the team would apply to any other engineering change. Pilot data, defect rates, supervisor feedback, and observed performance, rather than enthusiasm and the desire not to look unambitious.

This dimension is uncomfortable to score honestly because the pressure to expand agent scope usually comes from outside the team, often from people whose incentives are about progress narrative rather than delivery quality. A team scoring well here is a team that has learned to say not yet, and to back it with data when challenged.

Scoring rubric

LEVEL 1 Absent	Agent scope expands based on demos, excitement, and seniority of the requester. No data is consulted. Reversals are rare and reluctant.
LEVEL 2 Informal	Some informal review happens before expansion. The standard is mostly intuition. Decisions feel reasonable in the room but cannot be reconstructed afterwards.
LEVEL 3 Documented	An expansion process is documented. It exists. It is used selectively, often around the more visible decisions.
LEVEL 4 Complete	Expansion requires evidence. Pilot data, defect rates, and supervisor input are formally consulted. Decisions are traceable.
LEVEL 5 Substantive	Capability decisions are quantitative, reversible, and held to the same standards as engineering changes. Reversals happen when evidence does not support continuation. The team can name a recent expansion that was paused, narrowed, or rolled back.

What signals each end of the scale

SIGNALS YOU ARE AT LEVEL FIVE

- Expansion proposals are reviewed against pilot data, not stakeholder enthusiasm.
- An agent capability has been narrowed or paused based on evidence in the last quarter.
- The expansion process produces visible decisions, including refusals.
- Sponsors of expansion are required to commit to the rollback criteria up front.

SIGNALS YOU ARE NOT

- Every expansion proposal in the last six months has been approved.
- Pilot data is collected after expansion, not before.
- The pressure to expand comes from outside the team and routinely overrides local judgement.
- No one can remember the last time an expansion was paused or reversed.

One concrete action to move forward

Take the next expansion request and write a one-page memo: what evidence justifies the expansion, what success looks like, and what would trigger a rollback. If the memo cannot be written, the expansion is not yet ready for a decision.

Running the assessment session

A good CDI session takes ninety minutes and produces a Canvas the team can defend. A bad one takes thirty minutes and produces a Canvas that flatters the team. The difference is mostly in how the facilitator treats the level four to five jump.

Who should be in the room

Five to eight people, including at least one named supervisor, one engineer who works closely with agents, one person with delivery accountability, and one person who would be answerable for an incident. Avoid sessions of only senior managers, which produce optimistic scoring, and avoid sessions of only engineers, which produce pessimistic scoring without context.

The session shape

1 FRAME THE EXERCISE (10 MINUTES)

Read the foreword to the room. Establish that the goal is honest scoring, not flattering scoring, and that the most useful answer is often a lower score than the team initially proposes. Emphasise the level four to five distinction.

2 SCORE EACH DIMENSION (10 MINUTES PER DIMENSION)

Take each dimension in turn. Read the diagnostic question. Ask each participant to silently mark a score. Reveal scores together. Where scores differ by more than one level, discuss the gap. Land on a single score, with the evidence noted in the canvas.

3 IDENTIFY THE STRONGEST AND THE WEAKEST (10 MINUTES)

Once all six dimensions are scored, identify the most mature dimension and the largest gap. The most mature is not always the highest scoring; it is the one where the evidence for the score is strongest.

4 COMMIT TO A NINETY-DAY PRIORITY (10 MINUTES)

Choose one dimension as the focus for the next quarter. The choice should be a movement of one level on a single dimension, not aspiration across all six. Write the action in the priority box.

Facilitator notes

The facilitator's job is to challenge level four claims. When the team scores a dimension at four, the question is always the same. Can you name a specific occasion in the last quarter where this control changed an outcome. If the answer is no, the dimension is at three, not four. If the answer is yes, with detail, the dimension may be at five.

Resist the temptation to score everything at three. Three is the comfortable middle, and a session full of threes usually means the conversation has not happened yet.

Reading your results

The headline number on a CDI assessment is less informative than the shape of the result. A team scoring three across the board is in a different position from a team with two fives and four ones, and the conversation each team needs is different.

Patterns to recognise

THE FLAT PROFILE

All scores within one level of each other. This is most common in teams new to agent work. The advice is to identify which one dimension would unlock the most movement on the others, and concentrate there. Provenance Integrity is often the right starting point because it is upstream of every other dimension.

THE AUDIT-READY PROFILE

Fours across the board, with no fives. This is the governance theatre profile. The team can pass an audit and is exposed to incident risk. The action is to pick one dimension, run the level five test on it, and use the result as a forcing function for the rest.

THE SPLIT PROFILE

Strong on some dimensions, weak on others. The diagnostic question is whether the strengths and weaknesses are connected. Strong provenance with weak review substance, for example, means the team is collecting evidence it does not use. The fix is not better provenance. It is better review.

THE BRITTLE PROFILE

One or two fives, and scores below three on others. This is more common than it should be. It usually reflects a team where one engaged individual has driven a single dimension to maturity, while the surrounding system has not kept up. The action is to make the strong dimension less dependent on the individual.

What the headline number does tell you

An average score below three suggests the team should not be expanding agent scope. An average score above four suggests the team is at risk of governance theatre, and should run the level five test on at least one dimension before celebrating. The most useful range to be in, given honest scoring, is between three and four with at least one dimension at five.

CALIBRATION CHECK

If the assessment took less than thirty minutes and the room broadly agreed throughout, recalibrate. Honest assessments produce disagreement, and disagreement is the signal that the dimensions are doing their work. A frictionless assessment is usually a flattering one.

From canvas to ninety-day action

A CDI assessment produces a Canvas. A Canvas without a commitment is a wall decoration. The next ninety days should produce one observable change on one dimension, and the discipline of choosing one is more important than the ambition of the choice.

Choosing the focus dimension

The focus dimension for the next quarter should meet three criteria. It should be the largest gap relative to where the team needs to be, not absolutely. It should be a dimension where movement is feasible in ninety days, given current capacity. And it should be a dimension whose movement would unlock progress elsewhere.

Provenance Integrity often meets all three criteria for teams early in their journey, because it is foundational and because the work is mostly tooling. Review Substance is a stronger choice for teams already at level four on most dimensions, because the cultural work needed to reach five elsewhere depends on review being real.

The ninety-day action shape

A good ninety-day action is specific, observable, and would produce a defensible level upgrade if completed. Three examples, one for each kind of profile.

EXAMPLE: MOVING PROVENANCE INTEGRITY FROM TWO TO THREE

"By the end of the quarter, every agent contribution will be tagged at the time of commit, with the prompt retained in a queryable store. We will run a reconstruction exercise on a closed defect and document the time taken."

EXAMPLE: MOVING REVIEW SUBSTANCE FROM FOUR TO FIVE

"By the end of the quarter, the agent submission rejection rate will be tracked and discussed in retrospectives. Reviewers will have protected time blocked for review. The team will publish three examples of substantive rejections from the period."

EXAMPLE: MOVING CAPABILITY CALIBRATION FROM THREE TO FOUR

"The next two expansion proposals will be reviewed against a written evidence standard. Each proposal will include rollback criteria. Decisions and their evidence will be logged in a central register."

Tracking progress between assessments

The team does not need to re-run the full assessment to know whether it is making progress. A weekly check on the focus dimension is enough. Has the action moved forward this week. Is the evidence accumulating. Will the dimension defensibly support the new score by the end of the quarter.

Common failure modes

Most teams that run a CDI assessment will recognise themselves in at least one of the patterns below. Naming the pattern is usually the first step out of it.

The optimism tax

The team scores itself one level higher than the evidence supports on most dimensions, because the alternative feels like criticism. Symptom: the assessment took less than an hour and produced no surprises. Remedy: re-run the assessment with one external participant whose judgement the team respects.

The documentation reflex

The team responds to a low score by writing a new policy. Within a quarter the policy exists and the score is unchanged. Symptom: the team can describe the document but not the behaviour. Remedy: refuse to call a dimension upgraded until a specific behavioural example can be cited.

The hero pattern

One person on the team holds a dimension at five through personal effort, while the surrounding system has not adapted. Symptom: when that person is on holiday, the score collapses. Remedy: invest in making the dimension robust to absence before adding new dimensions.

The audit chase

The team treats the CDI as an audit framework and works to satisfy each level on paper. Within two assessments, the scores look excellent and the substantive position has not changed. Symptom: the team can explain the scoring rubric better than its own delivery practice. Remedy: stop using the Index for external reporting until the internal practice has caught up.

The big bang ambition

The team commits to moving four dimensions at once and moves none of them. Symptom: the next assessment looks identical to the last. Remedy: pick one dimension, ignore the others, and accept that this is not a slower path to the goal but the only path.

The score deflation

The team scores itself harshly to signal humility. The result is a Canvas that flatters the team's self-image but provides no useful direction. Symptom: every dimension is scored at one or two, with no specific evidence cited. Remedy: ask for evidence at every level, including the low ones.

Re-assessment cadence

The Index is designed to be run quarterly, with a deeper assessment annually. The cadence matters because the value of the Index is in the trend, not the snapshot.

Quarterly assessment

A ninety-minute session, focused on the dimensions that have moved or that the team is working on. The Canvas can be updated rather than rebuilt. The conversation focuses on whether the focus dimension from the previous quarter has moved as planned, and what the new focus should be.

Annual deep assessment

A half-day, with all six dimensions re-scored from scratch. The aim is to detect drift on dimensions the team has not been actively working on. Scores often decline year on year on dimensions that have been left alone, because the surrounding system changes faster than the dimension's evidence does.

Triggered assessment

An off-cycle assessment is appropriate after any agent-related incident, any significant change in agent scope, or any change in the team's composition that affects the supervisory relationships. A triggered assessment focuses only on the dimensions implicated by the trigger.

What to do with old Canvases

Keep them. The diagnostic value of a sequence of Canvases is greater than any single one. Patterns that take a year to reveal themselves only do so when the historical record is intact.

A NOTE ON EXTERNAL REPORTING

The Index is designed for internal use. External reporting tends to invite optimistic scoring, which destroys the instrument. If you need to report governance maturity externally, abstract from the Index rather than using it directly. The temptation to communicate a CDI score to a sponsor is the temptation to inflate it.

About the framework

The Cognitive Delivery Index is one instrument within the Cognitive Delivery Framework, a body of work concerned with redesigning the software development lifecycle for hybrid human and agent delivery.

The wider framework

The Framework is published as a six-paper series on Substack, of which this Index is the measurement instrument introduced in paper six. The earlier papers cover the developer progression from Dev 1.0 through Dev 4.0, the redesign of code review for agent participation, the governance principles underpinning hybrid delivery, the question of accountability when agents act, and the adoption tiers organisations move through as agent participation matures.

The book-length treatment is forthcoming, with publication targeted by the end of 2026. Readers interested in early access to material as it is developed can subscribe via the links at the foot of this page.

How to engage with the work

The Index is offered freely. Teams are encouraged to use it, adapt it to their context, and share what they find. The most useful feedback is the kind that surfaces a dimension the Index does not capture, or a level distinction that fails to discriminate in practice. Both kinds of feedback shape future versions.

If your organisation has run an assessment and would be willing to share findings, anonymised where appropriate, please get in touch directly. The empirical base for the Framework is built one team at a time.

Version history

This is version 1.0 of the companion guide, accompanying version 1.0 of the Canvas. Future versions will reflect feedback from teams using the Index in production and any refinements to the dimensions or the level descriptors.



Jason Cruz Falzon
Creator of the Cognitive Delivery Framework
jason@cognitivedelivery.co.uk
cognitivedelivery.co.uk / cognitivedelivery.substack.com

This document is a companion to The Cognitive Delivery Index Canvas, version 1.0.