CONSTRUCT & MEASUREMENT-METHOD PAPER

Scoring the Pushback, Not the Decision

A Deterministic Index of Challenge Intensity in an AI Cognitive Contradictor — the Δ-CSI (Cognitive Sovereignty Index, delta)

by

Francesco Saverio Canepa

Independent researcher, Rome

WORKING PAPER · VERSION 1.0 · 2026-07-23 · CC BY 4.0

Companion papers: *Cognitive Sovereignty* (osf.io/6n8tg) · *Calibrated Dissent* (osf.io/5qc8t, DOI 10.17605/OSF.IO/5QC8T).

CONTENTS

Contents

CONSTRUCT & MEASUREMENT-METHOD PAPER

Abstract

Decision-making under intensive AI assistance faces a documented paradox: more analysis does not reliably produce better judgment, because assistants optimized for user satisfaction tend to confirm rather than contest, inflating confidence and homogenizing reasoning across the organizations that share them. There is a structurally adversarial alternative: a *cognitive contradictor*, whose function is to challenge a decision rather than validate it. It raises an immediate measurement question. How does one quantify *how hard* a decision was contested, in a way that is reproducible in its scoring, comparable over time, and honest about what it does and does not establish?

This paper defines that quantity. It introduces the **Δ-CSI (Cognitive Sovereignty Index, delta)** as a construct — the *intensity of the structured challenge* applied to a single decision in a single analysis session — and specifies its operational definition.

The Δ-CSI is a deterministic index in [0, 100], computed from five weighted components: implicit-assumption count modulated by mean severity, falsification-test count, burden-of-proof-question count, cited-source count, and thesis-versus-counter-scenario divergence. Counts are capped, which precludes gaming by verbosity. Weights are constrained to sum to one hundred, interpretation bands are fixed, and the versioning scheme is explicit. Two limits remain, and are stated openly: the self-reported severity channel is gameable by construction (§7), and the count components register quantity, not argument quality.

The paper contributes three things. (i) The construct, delimited by four negations: challenge intensity is *not* decision quality, *not* the probability of success, *not* a verdict on the decision-maker, and *not* whether the decision was subsequently changed. (ii) The specification, together with its reproducibility guarantee, its fail-closed defenses against sycophantic output, and a register of its declared weaknesses. (iii) The comparability and calibration discipline: cohorts keyed by computation version, divergence method, and weights, with statistics held to descriptive-only below five recorded outcomes.

The paper makes no claim of empirical validity or effectiveness. It does not assert that a high Δ-CSI corresponds to better decisions — nor, conversely, to weaker ones — and it does not assert that the index predicts outcomes; it demonstrates construct definition and measurement method only.

Empirical validation is stated as a deferred programme, structured as a pre-registered evidence-accumulation roadmap. Its two stages are per-tenant ledger calibration and the public sealed M&A pilot of the companion *Calibrated Dissent*, in which the Δ-CSI enters strictly as non-scoring provenance and an exploratory covariate. We treat this honesty as a design requirement, not a caveat: an index of challenge intensity that overstated its own validation would be inconsistent with the very stance it exists to measure.

Keywords: decision quality; cognitive contradictor; adversarial AI; construct validity; challenge intensity; calibration; AI governance; measurement.

SECTION 1

Introduction

Leadership now has analytical assistance of a kind without precedent. AI systems synthesize evidence, build scenarios, and return arguments faster and in more finished form than any prior staff function.

From this it is inferred, almost always without checking, that decisions so assisted are better decisions. There is reason to doubt it.

The model of the world on which a consequential decision rests is seldom tested before it becomes action. More often it is silently confirmed. The disposition to seek evidence that corroborates a preferred hypothesis is among the most robust findings in the study of judgment under uncertainty (Kahneman, 2011). And the assistant most pleasant to consult is precisely the one that supplies that evidence on demand.

The mechanism is documented, not conjectured. Language models fine-tuned from human feedback exhibit *sycophancy*: a measurable tendency to conform to the user's stated view, and to revise correct answers toward the user's error when pressed (Sharma et al., 2023).

The behaviour compounds two older findings from the study of human–automation interaction. *Over-reliance* is the tendency to accept an automated recommendation without the scrutiny one would apply to a human proposal (Parasuraman & Manzey, 2010). *Deskilling* is the loss of competence that follows when its exercise is progressively delegated to a system that performs it (Bainbridge, 1983).

A fourth effect operates across organizations rather than within them, because they increasingly consult the same small number of models: *algorithmic monoculture*. Correlated tools induce correlated judgments, and the errors of the shared system cease to be independent across the institutions that rely on it (Kleinberg & Raghavan, 2021).

The paradox, stated precisely: abundant analysis can aggravate the failure it appears to remedy. The more fluent the confirmation, the less a decision is tested — and the more its untested premises are shared with everyone using the same tools.

The companion paper *Cognitive Sovereignty* (Canepa, 2026a) develops the construct that names what is eroded under these conditions: the collective capacity to hold judgment independent, calibrated, and falsifiable. It also identifies the mechanism that would protect it — a system whose function is to contest a decision rather than to assist its author's reasoning.

Following that work we call the mechanism a **cognitive contradictor**, and we retain the term throughout. Softer labels will not do: "critic", "advisor", "devil's advocate" admit by connotation exactly the optional, advisory posture the mechanism is designed to exclude. A contradictor's output *is* the challenge: contestation is its objective function, not an occasional by-product of an analysis oriented toward synthesis.

A mechanism defined by the intensity of the challenge it applies invites an obvious question, and it is a measurement question. If contestation is the point, one must be able to say *how much* of it occurred. Otherwise the claim that a given analysis was adversarial is unfalsifiable, and the distinction between a substantive challenge and a performance of one collapses.

The quantity is not decision quality: unobservable at the moment of decision, and confounded by luck thereafter. Nor is it forecast accuracy, which is a separate property, measured by separate means (Brier, 1950), and audited in the companion protocol (Canepa, 2026b). The quantity is the *intensity of the challenge itself*: how forcefully, and along how many independent dimensions, the contradictor pressed on the assumptions of the decision under examination.

This paper defines that quantity and specifies its measurement. We name the index the **Δ-CSI — the Cognitive Sovereignty Index, delta**. The name is deliberate, and it is the first thing a careful reader will interrogate (§3.4).

The contribution is threefold. First, the *construct*: challenge intensity, delimited from four adjacent notions with which it is routinely and consequentially confused (§3). Second, the *operational definition*: a deterministic index in [0, 100], with its five components, its anti-gaming and comparability rules, its reproducibility guarantee, and its fail-closed defenses, fixed in a versioned specification (§4–§7). Third, the *validation discipline*: a precise account of what the index does and does not establish, and a pre-registered roadmap along which empirical validity could be accumulated — never assumed (§6, §8).

The stance of the paper on its own validation is not modesty. It is a methodological commitment, and it is load-bearing.

A measurement method presented in fluent, confident prose is especially exposed to selective reporting and overclaiming. And a project whose central premise is that fluent confirmation is precisely the thing to distrust cannot exempt its own account from that suspicion.

We take the lesson literally. Every factual claim below is anchored to the specification, to the reference implementation, or to a cited source. Every property that is not yet demonstrated is marked as such. The boundary between what is defined, what is engineered, and what remains an empirical conjecture is drawn explicitly and kept visible.

A construct whose purpose is to measure the intensity of self-contestation cannot be introduced by a document that declines to contest itself. Section 7 therefore turns the method on its own index, and records without mitigation the weaknesses that survive.

The remainder of the paper proceeds as follows. Section 2 positions the index against the literatures it borrows from and the measures it must not be confused with. Section 3 defines the construct and its four negations. Section 4 gives the operational definition. Section 5 situates the scorer in the pipeline that produces its inputs, and states the provenance/scoring separation that keeps the index interpretable.

Section 6 describes persistence and the calibration discipline: the sense, precise and limited, in which the index is "calibrated". Section 7 is the self-contradiction — the declared

weaknesses. Section 8 states the limits and the pre-registered validation roadmap. Section 9 sets out the terms on which the specification is offered as an open standard. Section 10 concludes.

SECTION 2

Related work and positioning

The Δ-CSI sits at the intersection of four literatures and must be distinguished from a fifth family of measures with which it shares vocabulary but not object. We take these in turn, because the paper's contribution is legible only against what it is *not*.

Structured challenge in decision-making. That the quality of a consequential decision depends on the architecture of the process that produced it is an old and well-supported finding. It does not depend on the competence of the participants alone, nor on the quantity of information available.

Schwenk's (1990) meta-analysis of devil's advocacy and dialectical inquiry provided evidence that structured disagreement can improve decision quality relative to consensus-seeking, with effects contingent on how the procedure is implemented. And it showed, crucially, that the effect is a property of the *procedure* rather than of the individuals enacting it.

The cognitive contradictor inherits this lesson directly: the mandate to challenge is codified in the architecture of the system, not delegated to the vigilance or character of a human advocate who may tire, defer, or be outranked. The Δ-CSI is the instrument that makes the intensity of that codified challenge observable, and therefore auditable, session by session.

Adversarial and antagonistic AI. A recent line of work argues that the near-universal design default of agreeable, accommodating AI is neither inevitable nor always desirable, and explores deliberately antagonistic interaction as a design space (Cai, Arawjo, & Glassman, 2024). The contradictor is a governance-oriented instance of that stance, narrowed to a single function — contesting a decision — and bound to a fixed output contract.

But the same literature supplies the reason the challenge must be *measured* rather than merely *invoked*. Controlled studies of multi-agent debate find that adversarial framings do not reliably produce genuine disagreement. Debating agents frequently converge, echo a shared prior, or perform opposition without altering the substance of their reasoning (Wynn, Satija, & Hadfield, 2025; Wu, Li, & Li, 2025).

An index of challenge intensity is, in part, a defense against this failure mode. It is only a partial one, and the paper is explicit about the limit.

A staged opposition that is also *thin* — few assumptions, low declared severity, a counter-scenario close to the thesis — does score low. But another kind escapes the index: a performance that supplies the structural markers of a challenge without its substance, with many self-rated-severe assumptions, verbose but banal tests, and a lexically distant yet not substantively opposed counter-scenario.

The current construction does *not* distinguish that performance from a genuine challenge. Four of the five components register structure — three of them pure counts, plus the

structural divergence distance — rather than grade quality, leaving assumption severity as the sole quality signal (§3.3, §7.2).

The index raises the cost of the crudest performed opposition; it does not detect the sophisticated kind. That gap is one of the declared weaknesses of §7.2, not a property the index claims to have.

Calibration and forecasting. The forecasting literature supplies both a discipline the Δ-CSI adopts and a measure it must not be mistaken for.

Calibration, in the sense of Tetlock and Gardner (2015) and the scoring tradition that precedes them (Brier, 1950; Murphy, 1973), is the correspondence between stated confidence and empirical outcome frequency. It is a property of *predictions*, established by comparing probabilities against realized events. The Δ-CSI is not a forecast and has no Brier score: it makes no probabilistic claim about any outcome.

What the index borrows is the *epistemic humility* of that tradition — the refusal to assert predictive validity in advance of outcome data. And it borrows the machinery only downstream, in the calibration discipline of §6, where accumulated Δ-CSI scores are compared against recorded decision outcomes per tenant.

The companion protocol (Canepa, 2026b) is where the contradictor's *forecasts* are audited against outcomes under a pre-registered plan. There, the Δ-CSI enters explicitly as non-scoring provenance, never as a scored quantity. The separation is deliberate and is maintained throughout the corpus: this paper measures the challenge; the companion audits the prediction.

Construct validity. Because the Δ-CSI is a proposed measure of an unobservable property, it belongs to the tradition of construct validation in the measurement sciences (Cronbach & Meehl, 1955; Messick, 1995). That tradition supplies the paper's standard of adequacy and its principal caution.

Cronbach and Meehl require that a construct measure be embedded in a *nomological network* of theoretical relations, against which its behaviour can be tested. It is a requirement the corpus can at present only *sketch*, not satisfy. The construct paper (Canepa, 2026a) situates challenge intensity among the dimensions of cognitive sovereignty, this paper operationalizes it, and the companion protocol (Canepa, 2026b) specifies the empirical tests.

But the relations this paper can state today are, in honesty, mostly *discriminant* — what the index is *not* tied to (§3.2) — and definitional rather than demonstrated. The *convergent* tests that would confirm the construct are set out in the roadmap (§8.2) and not run here. They would be, for instance, agreement between the Δ-CSI and independent expert ratings of the same challenges, or a known-groups separation of genuine from deliberately performed challenges. We state the network as an aspiration with a deferred confirmatory programme, not as an achievement.

Messick (1995) insists that validity is a property of *inferences drawn from scores*, not of the instrument as such. That dictates the paper's most important restraint: we validate no inference here. We define the score and its arithmetic; which inferences the score will license is an empirical question, reserved for the roadmap of §8.

Measuring argument and deliberation quality. The Δ-CSI counts and weights dialectical moves — assumptions, falsification tests, burden-of-proof questions — and so has ancestry in two measurement traditions the paper must acknowledge, because they are where a referee will look first.

In deliberation research, the Discourse Quality Index (Steenbergen, Bächtiger, Spörndli, & Steiner, 2003) scores the quality of political deliberation along coded dimensions of justification and respect. In computational linguistics, argumentation mining and argument-quality assessment (Wachsmuth et al., 2017) build taxonomies for grading the cogency and reasonableness of arguments in natural language.

The Δ-CSI differs from both in object, and the difference is deliberate rather than an oversight. Those instruments grade the *quality* of argumentation, typically through human or trained-model coding. The Δ-CSI measures the *intensity* of contestation through mostly structural counts, precisely so that its quantitative part is reproducible without a coding scheme (§3.3, §4). It buys reproducibility at the cost of quality-sensitivity — a trade the argument-quality literature makes in the opposite direction — and §7.2 records the resulting limitation rather than papering over it.

Adjacent, too, is the decision-analysis tradition that locates the quality of a decision in its *process* rather than its outcome (Howard, 1968). The Δ-CSI instruments one facet of that process — the intensity of applied challenge — without claiming to measure decision quality itself (§3.2).

What the Δ-CSI is not: the organizational CSI. A measure with a nearly identical name appears in the surrounding literature on cognitive security, and the collision is consequential enough to require a paragraph rather than a footnote.

The *organizational CSI* was introduced in the "Cognitive Rearmament" programme (De Luca, Giamminola, & Pollicino, 2026). It indexes the cognitive posture of an entire organization along structural dimensions: its dependence on algorithmic systems, the degree to which those systems frame decisions before they are consciously made, and the organization's residual critical capacity. It is a framework-level construct, assessed over an institution.

The Δ-CSI is a different object entirely. It is deterministic where the organizational CSI is framework-based, it operates on a single decision where the organizational CSI operates on an organization, and it measures the intensity of one challenge where the organizational CSI measures a standing posture. The two indices share three letters and a lineage of concern; they do not share a level of analysis, a method of construction, or an object of measurement.

A conceptual bridge between the session level and the organizational level is plausible: decisions habitually subjected to intense contestation might, over time, strengthen an organization's cognitive posture. But no such bridge is formalized, no calibrated mapping between the two indices exists, and nothing in this paper should be read as though one did.

Where this paper writes "CSI" without qualification it refers to the organizational index of De Luca et al. (2026); the session-level index is always written "Δ-CSI".

The positioning can be summarized as a single discipline. The Δ-CSI borrows the *procedural* insight of the structured-challenge literature. It instantiates the *adversarial* stance while defending against its known failure to produce real disagreement. It adopts the *humility* of the calibration tradition without claiming its predictive machinery. And it submits to the standard of *construct validity* while explicitly deferring the validation of any inference. What remains is to define the construct precisely enough that these commitments have teeth.

SECTION 3

The construct: challenge intensity

3.1 DEFINITION

The Δ-CSI measures the **intensity of the structured challenge** that a cognitive contradictor applies to the assumptions of a single decision, in a single analysis session.

"Intensity" here is a compound of extent and force. Extent: how many independent points of contestation the challenge raised — assumptions surfaced, falsification tests proposed, burdens of proof assigned, sources adduced. Force: how far its counter-scenario departed from the thesis it opposed, weighted by how grave the surfaced assumptions were.

It is a property of the *challenge*, produced in one session, about one decision. It is not a property of the decision, of the decision's future, or of the person who made it.

This is a modest object by design: the index does not adjudicate the dispute it measures, it measures the dispute.

An analogy makes the register precise. A boxer preparing for a bout does not ask a sparring partner to predict the result; the request is to be tested. At the end of a round the boxer can say how hard it was — how many blows had to be parried, how often they were forced onto the back foot. That judgment measures the intensity of the session, and it forecasts nothing about the fight.

The Δ-CSI is that judgment, rendered as a number: how hard the round was between a decision and its contradictor. It does not say whether the decision will win.

3.2 THE FOUR NEGATIONS

The construct is defined as much by what it excludes as by what it asserts. Four exclusions are constitutive. Each corresponds to a category error that the index's name and numeric scale actively invite.

Not decision quality. A high Δ-CSI does not mean the decision is wrong; a low one does not mean it is right. The index scores the challenge, not the choice. A decision may survive an intense contradictor intact — a high score with a sound decision — or collapse under a mild one. The number does not distinguish these cases, because it is not about the decision at all.

Not the probability of success. No component of the calculation estimates how the decision will turn out. The index is silent on outcomes by construction: it reads the structure of the challenge, and the structure of a challenge contains no forecast. Forecast accuracy is a distinct property with its own instruments (Brier, 1950), audited elsewhere in the corpus (Canepa, 2026b).

Not a verdict on the decision-maker. The index is not a grade for the person or group that formed the decision, nor an assessment of the work done before the analysis. A poorly formed decision may attract a high score precisely because it offered the contradictor abundant purchase. A well-formed one may attract a low score because little remained to contest. The number describes the contest, not the contestant.

Not whether the decision changed. Most consequentially: the Δ-CSI does not measure whether the contradictor actually moved the decision-maker's judgment. That effect — did the analysis change the decision before it was enacted — is a separate, observable fact. The ledger records it separately, as a Boolean field (`decision_changed`), and never folds it into the index.

Challenge intensity and judgment change are distinct. An intense challenge may leave a decision unchanged: the decision-maker considered the objections and held. A mild one may change it: a single well-placed question sufficed. Conflating the two would make the index measure its own persuasiveness, which is not what it measures.

These are not caveats appended to a measure that would otherwise assert those things: they are the boundary of the construct. A category error is not corrected by more precision. Asking the Δ-CSI whether a decision is sound is like asking a thermometer whether the weather is pleasant, and no threshold on the thermometer answers the question. Keeping challenge intensity separate from decision quality, from prediction, from merit, and from persuasion is the condition under which the index is useful rather than misleading.

3.3 A PROXY, AND WHAT KIND

The Δ-CSI is a proxy: an indirect quantification of an unobservable property through observable signals. This is unremarkable in the measurement sciences, where most constructs of interest are latent and reached through proxies. But the *kind* of proxy matters. Construct validity (Cronbach & Meehl, 1955) requires us to state two consequences.

First: the index measures the *intensity of the challenge produced*, not the *quality of the arguments* within it.

Three of its five components are counts (§4). A falsification test contributes to the score whether it is incisive or banal. The scorer counts items the pipeline labels as falsification tests, and does not itself verify that each is a genuine, criterion-bearing test rather than an exhortation — a limitation inherited from the count-only design (§7.2).

The declared construct holds: if the object is the intensity of contestation, the volume of well-formed adversarial artifacts is a partial but operationally defined indicator. It does not hold as a measure of argumentative quality, which the index does not capture and does not claim to.

The single exception is the assumption component, modulated by the mean severity of the assumptions surfaced. Severity is the one quality signal that enters the number, and it enters through a channel with its own vulnerability, as §7 records.

Second: the proxy is *honest* in a specific, checkable sense. It is deterministic, decomposable, and versioned: any two people computing it from the same challenge obtain the same number, and can attribute that number to its parts.

Honesty of a proxy is not validity of the inferences drawn from it (Messick, 1995). The former we guarantee in §4, the latter we defer in §8. Calling the Δ-CSI "an honest proxy" is a claim about its construction, not about its predictive power.

3.4 THE NAME: WHAT THE DELTA MARKS

The index's name expands to *Cognitive Sovereignty Index, delta*. A paper that introduces the index cannot decline to say what the delta marks: the question is the first a reviewer will raise, and it is raised against the paper's own object. We fix the answer here as a naming convention, stated rather than derived.

The delta marks the level of analysis. It distinguishes a *session-level* quantity — the pressure applied within a single analysis of a single decision — from the *organizational-level* index: the CSI of De Luca et al. (2026), which assesses the standing cognitive posture of an institution.

The delta is not a temporal difference: it is not a change in the index between two times. And it is emphatically not a difference in the decision-maker's judgment, which is `decision_changed`, and which the index excludes (§3.2). It is a level marker. The Δ-CSI is to the organizational CSI as a single measurement is to a standing characteristic, and the delta records that the two operate at different levels with no formalized bridge between them (§2).

We adopt the convention explicitly because no prior document in the corpus defined the delta, and an index cannot leave its own name unexplained. Where a reader expects a difference operator, we ask them to read a level index instead. It is a convention of naming, and we mark it as such rather than dressing it as a result.

3.5 THE CONSTRUCT IN THE WIDER FRAME

Within the framework of *Cognitive Sovereignty* (Canepa, 2026a), the Δ-CSI is described as a *partial anchor* of the construct: one observable quantity among those by which cognitive sovereignty becomes measurable. It is not its overall measure, and not a predictor of the decision's correctness.

This paper is consistent with that placement and does not extend it. The Δ-CSI anchors one dimension — the observability of applied challenge — at one level, the session. It does not measure cognitive sovereignty: it measures one thing that a sovereign decision process does, once, and makes that one thing auditable.

SECTION 4

Operational definition

This section gives the index its exact form.

The definitions are normative. They are fixed, ahead of any further prose, in an accompanying specification: the `Δ-CSI Specification v1.0`, which binds computation version `csi-v1`. The account here is the readable statement of that contract, and where the two are read together the specification governs. The design intent is that the formulas below can be implemented and checked without recourse to any other source.

4.1 THE INPUT: A STRUCTURED CHALLENGE

The index is computed over a **challenge**: the structured product of the contradictor's analysis of one decision.

For scoring, a challenge reduces to four counts and one scalar. The counts are the number of implicit assumptions surfaced — each with an optional integer severity on a 1–5 scale — the number of falsification tests, the number of burden-of-proof questions, and the number of cited sources. The scalar is a single divergence value in [0, 1], derived from the thesis text and the counter-scenario text. The scorer consumes only these: the counts, the assumption severities, and the divergence scalar. It does not read the free text of the challenge beyond counting.

This boundary is what makes the quantitative part reproducible, while the narrative part — produced by a language model — is not (§5). It is also the precise sense in which the index measures the *structure* the pipeline produced rather than the *substance* of the arguments (§3.3).

4.2 THE FIVE COMPONENTS

The Δ-CSI is the sum of five independent contributions. In the formulas, `norm(n) = min(n, 5)/5` is the capped count normalizer, `severity_avg ∈ [1,5]` the mean assumption severity, `d ∈ [0,1]` the divergence; the core weights are (25, 25, 20, 10, 20).

COMPONENT	FORMULA	MAXIMUM
assumptions	`25 · norm(n) · (severity_avg / 5)`	25
falsification tests	`25 · norm(n)`	25
burden-of-proof questions	`20 · norm(n)`	20
cited sources	`10 · norm(n)`	10

COMPONENT	FORMULA	MAXIMUM
divergence	`20 · d`	20

The five sum to a score in [0, 100]. The components are additive with no cross-terms, so any score is decomposable: given a Δ-CSI, one can always state how many points came from each dimension. This is a deliberate property. An opaque aggregate would defeat the auditability that is the whole point of measuring the challenge.

Two modelling choices here are stipulations, not results, and the paper marks them as such.

The Δ-CSI is a *formative* composite: its components constitute the index by definition, rather than reflecting a single latent variable that causes them — as items on a reflective psychometric scale would. Two obligations follow, and neither is discharged in this paper.

The first concerns the selection of components. That it is exactly these five must be argued to span the domain of challenge intensity. It is a content-coverage claim the paper asserts but does not establish: the five in fact mirror the contradictor's fixed output schema.

The second concerns the form. The additive, compensatory form is itself an assumption. A linear sum lets a strong showing on four components buy back a weak one on the fifth; a conjunctive construct would instead let a weak channel cap the whole. Both the component selection and the additive form are listed among the declared weaknesses (§7.2).

This also qualifies the construct-validity framing of §2. A formative composite is validated primarily through content coverage and criterion relations, not through the internal-consistency logic of the reflective tradition that Cronbach and Meehl and Messick largely address. That is why the roadmap's convergent and known-groups tests (§8.2), rather than internal-consistency statistics, are the relevant evidence.

4.3 TWO RULES THAT KEEP THE NUMBER HONEST

Two construction rules do the load-bearing work, and each answers a specific way the number could be corrupted.

The cap, against inflation by volume. Every count-based component saturates at five: `norm(5) = norm(50) = 1.0`. The marginal value of a sixth assumption, or a tenth falsification test, is exactly zero. Ten shallow objections do not outscore five deep ones.

The rule removes the incentive to fabricate volume. It forces the route to a higher score through the substance of the challenge — a more distant counter-scenario, or assumptions rated graver — rather than the length of its lists. That second route, however, itself runs through the self-reported severity channel whose vulnerability §7.2 records. A reader should therefore never infer an exact count from a saturated component: a maximal `falsifications` contribution certifies "at least five", not a total.

The sum-to-one-hundred constraint, against covert inflation of the ceiling. The five weights sum to one hundred, so each weight is the maximum share its component can contribute, and the maximum possible score is one hundred.

The sum-to-one-hundred is itself a stipulated convention, enforced programmatically. What follows *from* it, given fixed weights, is that no reweighting can covertly inflate the ceiling.

A domain configuration (a "Pack") may redistribute the weights — a clinical or legal-diligence setting might raise the weight of sources at the expense of divergence — but only on the condition that the sum remains one hundred. The constraint forbids raising one component's ceiling without lowering another's. The interpretation bands (below) are never movable. The consequence, which §6 develops, is that scores produced under different weightings are not directly comparable and must be kept in separate calibration cohorts.

4.4 ASSUMPTION SEVERITY: THE ONE QUALITY CHANNEL

The assumption component is the only one modulated by a quality signal rather than a pure count. Its multiplier `severity_avg / 5` ranges from 0.2 (all assumptions at severity 1) to 1.0 (all at severity 5): two challenges with the same number of assumptions but opposite gravity produce different scores.

Two default conventions govern missing data. A single assumption with no declared severity counts as 3, a neutral midpoint. But a challenge that declares *no* severities at all is assigned `severity_avg = 5` — the maximum — for backward compatibility with the pre-severity computation version.

The second convention is a known defect, which we flag here and revisit in §7: it rewards omission over honest mid-range declaration. It is an inverted incentive, one that production pipelines are directed to avoid by always declaring severities, and that the paper declines to conceal.

4.5 DIVERGENCE: LEXICAL AND SEMANTIC

The divergence component measures how far the counter-scenario departs from the thesis, on a [0, 1] scale. It does so by one of two methods, which measure *different properties* and are therefore not mutually comparable.

The lexical method is the Jaccard complement over normalized token sets: `1 − |T(τ) ∩ T(κ)| / |T(τ) ∪ T(κ)|` (Jaccard, 1912). It is deterministic, offline, and dependency-free, and serves as the guaranteed fallback. It does, however, register a thesis and a counter-scenario that agree in substance but differ in wording as divergent — a distortion relative to the property of interest.

The semantic method maps the cosine similarity between embedding vectors of the two texts to a distance: `(1 − cos)/2`, in the vector-space tradition of Salton, Wong, and Yang (1975). It is more faithful to the construct, at the cost of a dependency on an external encoder.

Which method ran is selected by configuration and, critically, is *recorded on every score*. If a semantic encoder is requested but fails at runtime, the computation falls back to the lexi-

cal method and labels the record accordingly: scores of genuinely different kinds are never silently merged.

The two divergence measures are not interchangeable even at equal numeric value. The point recurs as a comparability constraint in §6.

4.6 INTERPRETATION BANDS AND VERSIONING

The score is read through three fixed bands — *marginal* [0, 33), *moderate* [33, 66), *substantial* [66, 100] — identical across every domain and configuration.

Fixity is deliberate. Were installations free to move the thresholds, the word "moderate" would lose its shared meaning, and the same number would tell different stories to different audiences. The canonical interpretation is always a statement about *pressure on the decision's assumptions* — "the contradictor applied moderate pressure" — never about the decision's merit or a change in its author's mind.

The computation is versioned. The current version (`csi-v1`) differs from its predecessor (`csi-v0`) in exactly one respect: the severity modulation of the assumption component.

Records carry their version. And because a severity-declared challenge scores differently under the two versions, scores from different versions are held in separate cohorts and never pooled (§6). Versioning is not bureaucratic overhead here: it is what allows the definition of the index to change without silently invalidating the scores computed under the prior definition.

4.7 A WORKED EXAMPLE

The construction is concrete on a single case, reproduced from the reference implementation and verifiable by hand.

A challenge surfaces three assumptions with no declared severities (hence `severity_avg = 5.0`), two falsification tests, two burden questions, one source, and a lexical divergence of 0.9615. The components are `25·(3/5)·(5/5) = 15.00`, `25·(2/5) = 10.00`, `20·(2/5) = 8.00`, `10·(1/5) = 2.00`, and `20·0.9615… = 19.23`, summing to **54.23** — in the *moderate* band. The number is produced deterministically by the reference scorer: given the same five inputs, it returns 54.23 every time.

This canonical example is retained for continuity with the reference documentation, but it deliberately runs on two of the index's weakest paths. We flag that rather than hide it. The assumptions carry *no declared severities*, so the assumption component runs at the backward-compatibility maximum (`severity_avg = 5.0` — the omission incentive of §4.4 and §7.2). And the divergence is the *lexical* Jaccard fallback, the less valid of the two methods (§4.5).

A methodologically cleaner illustration would declare mid-range severities and use semantic divergence. The number would differ, but the point of the worked example — the arithmetic and its by-hand reproducibility — would be unchanged.

The narrative that a real challenge would wrap around these inputs — the actual text of the assumptions, the scenario, the questions — is generated by a language model and is *not* deterministic. Only the quantitative part reproduces exactly. That distinction, between a reproducible number and a non-reproducible text, is the subject of the next section.

SECTION 5

The measurement pipeline and the provenance/scoring separation

The Δ-CSI does not arise from nowhere. It is the last link in a chain, and both its guarantees and its limits are legible only against the process that produces its inputs.

The contradictor operates as three sequential, isolated passes, orchestrated deterministically. The first (Thesis) extracts the decision's dominant reading — what a competent analyst would conclude — and is not itself adversarial.

The second (Red Team) is the challenge proper. It surfaces the implicit assumptions, constructs the counter-scenario, proposes the falsification tests, and poses the burden-of-proof questions. It draws its mandate, its bias taxonomy, and its falsification templates from the resolved domain configuration.

The third (Synthesis) calibrates confidence in light of thesis and challenge, declares the residual gaps, and attaches sources with their provenance. It has no mandate to conclude or to recommend.

Two features of the pipeline bear directly on the meaning of the index.

The first is *model isolation across passes*. The passes may run on different model providers, so that a decision's dominant reading is not attacked by the same model that produced it. The providers actually used are recorded with the output.

The motivation is a failure mode documented in the multi-agent-debate literature (Wynn, Satija, & Hadfield, 2025; Wu, Li, & Li, 2025): adversarial framings imposed on a single model tend to produce correlated reasoning and performed rather than genuine disagreement. Isolating the passes across providers is a structural attempt to de-correlate systematic blind spots, so that the challenge the Δ-CSI measures is more likely to be real.

It is a mitigation, not a proof. Whether it succeeds is one of the things an eventual validation would have to establish.

It also carries a cost the paper does not hide. Running passes on different providers injects between-model variance into the very counts and severities the index is computed from: the same decision may yield different components — and therefore a different Δ-CSI — under different provider routings. The reliability consequence is taken up below and listed as a weakness (§7.2).

The second feature, and the more important for the index's interpretability, is the **separation of provenance from scoring**. The companion construct paper treats it as one of the contradictor's four invariants (Canepa, 2026a, §4.4); we restate it here in its measurement form.

A contradictor equipped with retrieval could introduce a circularity. Evidence found during analysis might simultaneously shape the content of the challenge *and* inflate its

measured intensity. A high Δ-CSI would then become an artifact of how many sources happened to be available rather than of how forcefully the reasoning was contested, and the index would cease to measure the property it names.

The separation forecloses this. Sources are retrieved and recorded as provenance, but the Δ-CSI's components are computed from the *logical structure* of the challenge — assumptions, falsification tests, burden questions, scenario divergence — not from the volume or quality of retrieved material. The source count enters as one separate component, with its own bounded weight (ten of one hundred), and does not retroact on the others. This is the condition under which the Δ-CSI remains interpretable as challenge intensity, rather than a disguised measure of retrieval richness. The four non-source components are insulated from the information context, and the source component is bounded at ten of one hundred: retrieval richness can nudge the score, but cannot dominate it.

An index conflated with the abundance of sources could not be calibrated in the forecasting sense at all (Tetlock & Gardner, 2015), because it would be measuring a property of the corpus rather than of the reasoning.

The pipeline thus produces two qualitatively different outputs, and the distinction is essential to what the paper claims.

The narrative — thesis, scenario, assumptions, questions — is generated by a language model and is not reproducible: the same decision analysed twice yields different texts. The score is computed by a deterministic function over the extracted structure and *is* reproducible: the same components always yield the same number.

When the paper claims reproducibility (§4.7, §7), it claims it for the scorer, not for the pipeline. A regression test of the index must fix the components as inputs, and must not expect determinism from the narrative stage. Confusing the two would be a claim the system cannot honor, and we do not make it.

The honest consequence is that the *measure*, end to end, is not reproducible in the way the *scorer* is. Because the counts, severities, and divergence are extracted from a stochastic model, the same decision analysed twice will in general receive a different Δ-CSI. The determinism the paper guarantees is that of the aggregation given fixed components: a property of the arithmetic, not of the measurement.

How much the end-to-end score varies across runs and across providers — its test–retest and inter-model reliability — is unquantified here. It heads the validation roadmap of §8.2 and is among the declared weaknesses of §7.2.

SECTION 6

Persistence and the calibration discipline

An index computed once and discarded measures a moment. An index persisted and revisited can, in principle, be validated.

The Δ-CSI is written, for every session, to a per-tenant, local-first **Sovereign Decision Ledger**. The ledger records the decision, the full challenge, the index with its component breakdown and version, and — in deferral, when it becomes known — the outcome. The outcome has two fields: an `outcome_label` (correct, wrong, mixed, or unknown) and the Boolean `decision_changed`.

The ledger is per-tenant and isolated, and its accumulated record of challenges and outcomes is the substrate on which calibration operates. For the purposes of this paper it matters not as a commercial asset, but as the mechanism that makes empirical validity *constructible over time* rather than assumable in advance.

It is necessary to be exact about what "calibrated" means for the Δ-CSI: the word is doing narrower work than the forecasting literature from which it is borrowed would suggest. Calibration here is a *discipline*, not a state the index has attained. It has three rules.

Cohorts. Calibration statistics are computed per cohort, where a cohort is the triple (computation version, divergence method, weights).

Scores produced under different versions, different divergence methods, or different weightings are not comparable, and are never pooled. A Δ-CSI of 54 under lexical divergence and a Δ-CSI of 54 under semantic divergence do not assert the same thing about challenge intensity: their numerical identity is accidental (§4.5). Mixing them would produce statistics drawn from incommensurable measurements, which cannot be interpreted as samples of one variable.

A descriptive-only floor. Below a minimum of five outcome-bearing records in a cohort, the calibration output is explicitly flagged as descriptive only, and no conclusion about the index's validity is permitted.

This is not a significance threshold: it is an operational floor of minimum good sense. Clearing it licenses no validity claim, and even above it the evidence remains descriptive and local to the tenant. The gate exists to prevent a specific error: presenting the Δ-CSI as empirically validated before any outcomes have accrued.

A local, descriptive statistic. Above the floor, the calibration may report a `hit_rate`: the proportion of *changed* decisions with a definitive outcome that turned out correct.

The definition is deliberately narrow. It conditions on `decision_changed`, because the question calibration can honestly ask of the ledger is whether the contradictor, *when it actually moved a decision*, moved it well.

The `hit_rate` may be undefined even above the floor: if no decision in the cohort was ever changed, the denominator is zero. This is not an anomaly, but a correct report of the

absence of the relevant evidence.

Where defined, the `hit_rate` is per-tenant, descriptive, and local. It is not a causal estimate, and a high value in one domain implies nothing about another.

Two boundaries of the calibration apparatus deserve statement, because their omission would overstate what the ledger delivers.

First, the apparatus as specified does not relate the *value* of the score to outcomes. None of its statistics, `hit_rate` included, is a function of the Δ-CSI magnitude: the apparatus accumulates the evidence a future validity analysis would require, without itself constituting a test of the index's validity.

Second, the engine-level floor of five outcomes is far weaker than the statistical power a confirmatory study would demand. The companion protocol (Canepa, 2026b) sets that bar explicitly: a floor of at least twenty scored cases per horizon, for any confirmatory calibration test.

It is important that the two thresholds not be confused. Clearing the product's five-outcome gate authorizes a tenant to *read* descriptive statistics, not to *claim* validation. The distance between those two acts is the distance between this paper and its empirical successor.

At the time of writing, that distance has not been closed. No per-tenant cohort has reached even the five-outcome descriptive floor with real decision-makers: the flywheel is specified and instrumented, but not yet turned by live use.

The paper reports this as a fact about the state of the evidence, in the same register the companion protocol uses to report the pre-outcome status of its sealed pilot. The honest disclosure is part of the contribution: a measurement paper that obscured the emptiness of its own outcome ledger would be committing, in miniature, the overclaim the whole programme exists to resist.

SECTION 7

Self-contradiction: properties, defenses, and declared weaknesses

An index whose purpose is to measure the intensity of self-contestation must be willing to contest itself, and the willingness has to be demonstrated rather than asserted.

This section turns the method on its own object. It states first the structural defenses that make the index fail visibly rather than flatter silently; then, without mitigation, the weaknesses that survive those defenses.

7.1 THREE FAIL-CLOSED DEFENSES

The system implements three structural protections against sycophantic output, structural in the sense that they are code-level controls that fire independently of what the model produces.

The first is the *contradiction floor*. A challenge is admissible only if it contains at least one assumption, at least one falsification test, at least one burden question, and a non-empty counter-scenario.

If the floor is unmet, the system performs exactly one sharpened retry. If it is still unmet, the pipeline raises a controlled error and emits nothing. When there is no genuine challenge to be had, the system fails visibly rather than degrading toward a reassuring report: an empty challenge becomes an error, not a result.

The floor is a floor and not a ceiling, and its threshold is deliberately low — which is itself one of the declared weaknesses below. But its function is categorical, and it is the first line against the failure mode the whole index exists to detect.

The second is *closed-enumeration provenance*. Every source must declare a provenance from a closed set of five values. With no retrieval available, the synthesis is constrained to return an empty source list rather than inventing citations.

A source cannot materialize from nowhere: it has a declared origin, or it does not enter the count. This closes the most insidious failure of a system that cites, namely the fabrication of plausible but non-existent references.

The third is *honest sub-threshold calibration* (§6): below five outcomes the system declares its statistics insufficient, rather than simulating a validation it does not have. The defense is again one of visible failure over silent degradation. The correct output of an empty ledger is a declaration of insufficiency, not a reassuring number.

7.2 DECLARED WEAKNESSES

The following weaknesses are open. They are stated here because a measurement whose validity rests partly on the honesty of its self-account cannot treat its own limitations as a threat to be managed. In a due-diligence reading, the value of the list is not the absence of entries: it is the fact that the system declares them itself.

The name and scale invite a category error. The label "Cognitive Sovereignty Index" and the 0–100 scale invite reading the number as a measure of correctness. It is not (§3.2).

The risk is communicative, and is not neutralized by the current implementation. The mitigation is definitional discipline — the index must always travel with an explicit statement of what it measures — not a structural guarantee. This is the paper's own first-listed risk, and it is the reason §3 precedes §4.

Three components measure only quantity. Falsification tests, burden questions, and sources are pure counts (§4): five trivial elements score as five incisive ones, and argumentative quality does not enter the number. Only the assumption component carries a quality signal.

The floor is permissive. The contradiction floor requires one element per field: a challenge with a single obvious assumption, one banal test, one question, and a one-line scenario clears it. The defense against empty challenges is real, but its threshold is low. A thin-but-non-empty challenge passes and presents with a marginal score — which is to be read with suspicion (§3.2), not as reassurance.

A sourceless synthesis passes without warning. An empty source list clears the floor and the schema validation. The synthesis is stored and scored with the source component contributing zero, and no explicit signal flags the absence to the reader. The behaviour is defensible — an honest absence is better than a fabricated citation — but the silent zero is a declared gap.

Below five outcomes, nothing is empirically validated. Until a cohort accrues five definitive outcomes, calibration is descriptive and no validity conclusion is licensed (§6). This is a limit of evidence, restated as a weakness because it is routinely forgotten in the presentation of new metrics.

The one quality signal is self-reported by the model. Assumption severity — the sole quality channel that enters the number — is declared by the same model that generates the challenge.

The anti-inflation cap governs counts, not severities. A model that systematically overstates gravity inflates the heaviest component, and no structural defense intercepts it.

This is the index's most consequential open weakness, and it has the structure of a classic measurement pathology. Once a quantity computed from a model's own severity declarations is used to judge that model's output, the declaration is under pressure to serve the score rather than to describe the assumption. It is the collapse of a statistical regularity under control, in Goodhart's (1975) original formulation; and, in its now-standard generalization, the tendency of a measure to cease being a good measure once it becomes a target (Strathern, 1997, following Hoskin).

The index is, in this respect, gameable by construction through the severity channel, and it says so. The vulnerability is also a specific case of a documented general one. Using a language model to evaluate language-model output imports the documented biases of the LLM-as-judge setting into the single channel that carries quality information — above all self-preference, a model scoring its own output favourably (Zheng et al., 2023).

Omitting severities scores higher than declaring them honestly. For backward compatibility, a challenge that declares no severities is assigned the maximum mean severity. An all-omitted challenge therefore outscores an identical one with honestly declared mid-range severities (§4.4).

It is a known perverse incentive, accepted as the cost of version compatibility. Production pipelines are instructed to always declare severities; the defect is disclosed rather than repaired.

The end-to-end measure has unquantified reliability. The determinism the paper guarantees is the scorer's, given fixed components. The components themselves are extracted by a stochastic model: the same decision re-analysed will in general receive a different Δ-CSI, and the model-isolation routing of §5 adds between-provider variance on top.

The test–retest and inter-model reliability of the end-to-end score — and hence the stability of the band a decision lands in — is not measured in this paper. Because reliability is prior to validity, this is the most consequential of the pre-empirical gaps, and it heads the roadmap of §8.2.

The weights and the additive form are stipulated, not derived. The core weights (25/25/20/10/20) are a default judgment about the relative contribution of the five dimensions, not a quantity estimated from data or elicited from experts. And the additive, compensatory aggregation is a modelling assumption, not a demonstrated property of the construct (§4.2). A weight-sensitivity analysis — how often band assignments flip under reasonable reweightings — would bound the arbitrariness, and is not yet run.

The assumption component is not monotone in genuine challenge. Because the component multiplies a capped count by the *mean* severity, adding a further genuine but milder assumption can *lower* the score. Five assumptions at severity 5 yield the maximum 25.00, but a sixth genuine assumption at severity 1 drops the component to 21.67: the count is already saturated, and the mean falls.

A strictly larger, honest challenge can thus score lower. The non-monotonicity also creates an incentive to *suppress* mild assumptions in order to keep the mean high — orthogonal to the verbosity cap.

It is a real property of the current `csi-v1` formula, disclosed here rather than silently carried. A monotone aggregation, for instance a severity-weighted sum under the same cap, is a candidate correction for a future computation version.

Under a real encoder, semantic divergence cannot reach its ceiling. The cosine mapping `(1 - cos)/2` assumes cosine spans [−1, 1]. But sentence-embedding cosines for natural-language pairs are typically positive and high: the semantic divergence occupies roughly the lower part of [0, 1], and its component realistically contributes only a few of its twenty points. The

fixed bands therefore partition structurally different effective ranges across the lexical and semantic cohorts, and the same challenge scored under the two methods can land in different bands.

Cohorting keeps the two out of one calibration pool (§6), but it does not make the semantic scale reach as high as the lexical one. A distribution-calibrated mapping of the cosine would be the principled fix.

Several of these entries — semantic divergence as an available default, severity modulation of the assumption component — were themselves the resolution of earlier entries in this same self-contradiction. Others — the reliability gap, the non-monotonicity, the stipulated weights — were surfaced by turning an adversarial review on this paper itself.

That is the intended dynamic: the register of weaknesses is a live backlog, not a confession appended once. What the list demonstrates is not that the index is free of defects, because it is not. It demonstrates that its defects are surfaced by the same adversarial discipline the index applies to the decisions it scores. A vendor presenting these as already solved would be asserting more than the system asserts of itself.

SECTION 8

Limits and a pre-registered validation roadmap

The paper's claims are bounded, and the boundary is the point.

What is established here is a construct and a measurement method: the Δ-CSI is defined, its arithmetic is deterministic and reproducible, its defenses and weaknesses are specified, and its comparability rules are fixed.

What is *not* established is any empirical validity — that a given Δ-CSI level corresponds to anything about the decisions it scores. That claim is deferred, and this section states the terms of the deferral precisely enough to be held to them.

8.1 WHAT THE INDEX IS NOT: CORROBORATION

A reader trained in the philosophy of science will hear, in a phrase like *the intensity with which a decision was contested*, an echo of Popperian corroboration and of the severe-testing tradition that formalizes it (Popper, 1959; Mayo, 1996).

The echo is worth addressing directly, because the resemblance is partly real and partly a trap. It is real in spirit: both traditions prize confronting a claim with what could overturn it, rather than with what would comfort it.

It is a trap if read as identity, and the reason is precise. Severe testing is defined by the *probative power* of a test: a test is severe to the degree that it would probably have exposed an error had one been present. The Δ-CSI measures no such thing. As §3.3 and §7.2 state plainly, it counts falsification tests without grading whether any is probative: it captures the *quantity and intensity of contestation applied*, not the *severity* of any test in Mayo's sense. The word "severity" inside the index refers, moreover, to something else again — the self-rated gravity of an *assumption* (§4.4) — and should not be conflated with test-severity.

The decisive distinction, however, is temporal, and it is what makes "not corroboration" true rather than merely asserted. Corroboration scores a hypothesis by the tests it has *survived*, after outcomes are in. The Δ-CSI scores a decision by the challenges *applied to it*, before any outcome is known and regardless of whether the decision is later borne out.

The index is therefore not a corroboration measure under a new name. It is an audit of contestation applied ex ante, and its own validation (below) runs through decision outcomes in a ledger, not through the survival of hypotheses under experiment. It borrows the adversarial spirit of the severe-testing tradition; it does not borrow, and does not claim, its measure of probative severity.

8.2 THE EVIDENCE-ACCUMULATION ROADMAP

Validation is treated as a matter of process design over time, pre-registered rather than asserted, and it proceeds through thresholds that must not be conflated.

Before any of the outcome-dependent thresholds below, there is measurement work that requires no outcomes at all, and this paper defers it rather than performs it. Two studies are pre-registered here as the immediate next step.

The first is a *reliability study*: a fixed set of decisions run repeatedly, across at least two provider routings. It would report the run-to-run and between-model variance of each component and of the total score, a test–retest reliability coefficient, agreement on band assignment, and a standard error of measurement to attach to any reported Δ-CSI.

The second is a *discrimination study*: a labelled corpus of genuine challenges versus deliberately performed ("sham") ones. It is a known-groups test of whether the index separates them — the very capability §2 argues the index is *for*, and the one not yet shown.

A convergent check belongs to the same stage and requires no outcomes either: agreement between the Δ-CSI and independent expert ratings of challenge intensity, on the same challenges. It would give the otherwise discriminant-only nomological network (§2) its first confirmatory node.

None of these requires the ledger to fill, nor live users, nor decision outcomes. All three are the precondition, not the consequence, of the outcome work that follows, and their absence is precisely why this paper is a definition-and-method contribution rather than a validated measurement.

The first threshold is the engine's descriptive floor: five outcome-bearing records in a cohort, at which a tenant may begin to *read* the accumulating record of scores and outcomes descriptively (§6).

A pragmatic reading — enough to detect a signal worth pursuing — requires on the order of twenty to thirty records with a defined `decision_changed` field. That is a scale reachable by a single serious user over months, not a research programme.

Investor-grade evidence, in the corpus's own internal targets, is a different order of magnitude again: hundreds of definitive outcomes per vertical. It is named here only to mark the gap between a pilot's descriptive signal and a defensible empirical claim, so that the former is never presented as the latter.

The second threshold belongs to the companion protocol and validates an adjacent quantity, which the paper is careful not to over-attribute to the index.

Calibrated Dissent (Canepa, 2026b) instantiates a pre-registered, anti-hindsight audit. Its object is a fixed, publicly sealed pilot of ten high-stakes European M&A decisions, sealed at a single date (20 June 2026), each carrying forecasts resolved at two horizons: six to twelve months, and eighteen to thirty-six months.

That protocol scores the sealed *forecasts* against outcomes by proper scoring rules. It sets a confirmatory power floor of at least twenty scored cases per horizon — which the fixed pilot of ten is, by design, below. Two points of discipline follow, and both are load-bearing for the honesty of this paper.

First, in that protocol the Δ-CSI enters strictly as *non-scoring provenance* and, at most, an exploratory covariate. It is recorded, not scored, and no confirmatory hypothesis is stated on it.

Second, and consequently, the powered stage of the companion validates the contradictor's *forecasts*, not the Δ-CSI as such. A well-calibrated forecaster is evidence about the pipeline, not a direct validation of the challenge-intensity index.

The index's own validation route remains the per-tenant ledger calibration of §6 — the accumulated relationship between challenge intensity and recorded decision outcomes. That route, at the time of writing, has produced no outcomes yet.

8.3 CONFLICT OF INTEREST AND THE GAMING FAILURE MODE

Two disclosures close the section, in the register the companion protocol adopts for the same purposes.

The first is a conflict of interest, declared rather than managed away. The author of this paper is the author of the construct paper, the builder of the reference implementation, and a party with an interest in the measurement method being credible. A reader should weigh the paper's claims in that light.

The mitigations are the ones available to a single author: an open and irrevocable specification (§9), pre-registration of the empirical tests in the companion, and the intention to place outcome scoring under independent resolution. None of these removes the conflict. They make it visible and auditable, which is the most an honest disclosure can do.

The second is the index's own principal failure mode, already stated as a weakness (§7.2) and repeated here as a pre-registered falsification target: *gaming*.

Suppose a deployment were found to produce inflated Δ-CSI scores without any corresponding increase in the substance of the challenge, most plausibly through the self-reported severity channel. That would be evidence that the index had become a target and ceased to be a good measure (Strathern, 1997). The condition is stated in advance, as the discipline of pre-registration requires, so that its later appearance would count as disconfirming rather than being explained away.

SECTION 9

The specification as an open standard

The definition set out in this paper is offered under open terms, so that others may adopt it independently of the author's own implementation. This section records those terms briefly; they are governance rather than method, and nothing in §3–§8 depends on them. Four commitments fix them.

An open, irrevocable specification. The Δ-CSI specification is published under an open licence (CC BY), as an open and irrevocable definition rather than a permission that could later be withdrawn. Anyone may compute and use the index under the same public terms.

Arm's-length parity. The reference implementation is *one* conforming implementation, not the standard. It computes the Δ-CSI under the same public definition available to any other implementer, and it enjoys no privileged variant. The standard's authority rests on the definition, not on the implementation.

A citable identity of its own. The specification has an identity — a name, a document, a DOI — distinct from the product documentation of any implementation. What others cite when they cite the Δ-CSI is therefore the standard, not a vendor's manual.

This paper is that citable definition, and it supersedes prior internal descriptions of the index on the question of the definition itself. Implementation-internal documents remain authoritative on implementation behaviour, and subordinate to this specification on the definition.

A commitment to independent stewardship. Custody of the standard at this stage rests with the author as a natural person. That is sufficient for a definition whose priority and authorship are established by publication.

The commitment recorded here is that the name and specification are to be conferred on an independent steward, at the point where external adoption makes that separation necessary. Independent means structurally separate, in control and in funding, from any single implementer. The purpose is stated plainly: an index of independent judgment cannot credibly be governed, in the long run, by a party that profits from the scores. Placing the standard on open and irrevocable terms now is what keeps that future option open rather than foreclosed.

These commitments are governance, not mathematics, and they change nothing in §4. They are included because of what the alternative would be: a measurement standard whose canonical source is an internal product document, revisable at the vendor's discretion. That would undercut, at the level of stewardship, the independence the index exists to measure at the level of decisions.

SECTION 10

Conclusion

This paper has defined the Δ-CSI — the Cognitive Sovereignty Index, delta — as the intensity of the structured challenge that a cognitive contradictor applies to a single decision. And it has specified how that intensity is computed: deterministically, from five bounded and decomposable components, under anti-gaming and comparability rules fixed in a versioned specification.

It has drawn the construct's boundary in four negations: the index is not decision quality, not the probability of success, not a verdict on the decision-maker, and not whether the decision changed. Those four confusions are the ones the index's name and scale actively invite, and a measure of challenge intensity that permitted them would measure nothing usable. And it has been explicit, throughout, about the line between what is defined, what is engineered, and what remains to be shown.

What exists as of this writing is a definition, a reproducible method, and a discipline. A number that any two parties can compute identically from the same challenge and attribute to its parts. Defenses that make the absence of a real challenge fail visibly rather than flatter silently. A register of weaknesses surfaced by the same adversarial method the index applies to the decisions it scores.

What does not yet exist is empirical validity — evidence that a given challenge intensity corresponds to anything about the decisions scored — and the paper has neither claimed it nor obscured its absence.

Validation is set out as a pre-registered programme: per-tenant ledger calibration, and the sealed forecasting pilot of the companion, in which the index enters as provenance and covariate, never as a scored quantity. The empty outcome ledger is reported as the fact it is.

The stance is not incidental to the subject. A construct built to measure how honestly a decision confronts what it would rather not examine cannot be introduced by a document that spares itself the same treatment.

The Δ-CSI's first and most exacting test is the one this paper has tried to pass: to state a measure precisely, to defend it where it can be defended, to contradict it where it cannot. And to let the number stand only for what it actually measures, which is the intensity of the challenge, and not the quality of the decision.

APPENDIX

Δ-CSI Specification v1.0

This appendix is the normative specification the paper refers to in §4 and §9. Its own cross-references use the form A.N (e.g. A.9.1); references to §N without the A. prefix point to the main paper. It is frozen; any change requires a version bump.

Cognitive Sovereignty Index, delta — normative specification of the challenge-intensity index

ATTRIBUTE	VALUE
Specification version	`spec-1.0`
Binds computation version	`csi-v1`
Status	Normative reference. Frozen. Any change requires a version bump.
Language	English
Relationship to implementation	This specification defines the *standard*. `CounterBrain` (`brain_engine`) is the reference implementation; where the implementation and this document disagree, this document defines the standard and the implementation is non-conforming until reconciled. The implementation's internal source-of-truth (`METHODOLOGY.md`) documents implementation behaviour and is subordinate to this specification on the *definition* of the index.

> **Purpose.** This document fixes the definition, computation, invariants, and comparability rules of the Δ-CSI before any prose is written about it (contracts-first). It is self-contained: the formulas below can be implemented and checked without any other source. The worked example in A.11 is reproducible by hand and by the reference scorer.

A.1 — SCOPE AND READING ORDER

This specification defines a single scalar index, the Δ-CSI, computed over the structured output of one adversarial ("cognitive contradictor") analysis of one decision. It defines:

1. the objects the index is computed from (A.3);
2. the five component functions and their aggregation (A.4–A.7);
3. the interpretation bands (A.8);
4. the two divergence methods (A.9) and the specification's versioning (A.13);
5. the invariants a conforming implementation must not break (A.10);

6. a reproducible worked example (A.11);
7. the comparability (cohort) and calibration rules (A.12).

The Δ-CSI is an index of **challenge intensity**. It is not defined here, and must not be presented, as a measure of decision quality, of the probability that the decision succeeds, of the decision-maker's competence, or of whether the decision was subsequently changed. The interpretive burden of that distinction is discharged in the paper, not in this specification; this document only fixes the arithmetic and the contract.

A.2 — NOTATION

- `⌊·⌋`, `min`, `max` have their usual meaning.
- Counts are non-negative integers; scores are real numbers rounded as specified in A.7.
- A *token set* of a text is the set of its whitespace-separated, lowercased tokens (the reference tokenizer; A.9.1).
- `[a, b)` denotes a half-open interval, `a` included, `b` excluded.

A.3 — INPUT OBJECT: THE CHALLENGE

The Δ-CSI is computed from a **challenge** `C`, the structured product of the contradictor pipeline for one decision. For the purpose of scoring, `C` is the tuple:

`C = (A, F, B, S, τ, κ)`

where

SYMBOL	MEANING	CONSTRAINT
A	list of implicit assumptions, each carrying an optional integer severity $s_i \in \{1,\dots,5\}$	$n_a = \lvert A \rvert$
F	list of falsification tests	$n\_F = \lvert F \rvert$
B	list of burden-of-proof questions	$n\_B = \lvert B \rvert$
S	list of cited sources, each with a declared provenance (A.10.3)	$n\_S = \lvert S \rvert$
τ	thesis text (the dominant reading)	non-empty
κ	counter-intuitive scenario text	non-empty (A.10.2)

The scorer consumes only the counts n_a, `n_F`, `n_B`, `n_S`, the severity values of `A`, and a single divergence scalar `d` derived from `(τ, κ)` (A.9). It does not read the free text of `A`, `F`, `B`, `S` beyond counting and (for `A`) severity. This is a deliberate boundary: the index measures the *structure* the pipeline produced, not the intrinsic quality of the arguments (a limitation stated openly in the paper, not hidden here).

A.4 — COUNT NORMALIZATION AND THE CAP

Let `CAP = 5`. For any count `n`:

`norm(n) = min(n, CAP) / CAP`

so `norm(0) = 0`, `norm(1) = 0.2`, ..., `norm(5) = 1.0`, and `norm(n) = 1.0` for all `n ≥ 5`.

The cap makes the marginal contribution of every count element beyond the fifth exactly zero. It is the anti-inflation rule: verbosity cannot raise the score. A consequence for readers: a component at its maximum certifies "at least five elements", not an exact count.

A.5 — ASSUMPTION SEVERITY

The assumption component is the only one modulated by a quality signal. Define:

`severity_avg = (1 / n_a) · Σ_i s_i if A is non-empty and severities are declared`

with two default conventions:

- **Missing severity on a single assumption** counts as `3` (neutral midpoint), so an accidental omission does not penalize the challenge.
- **No severity declared anywhere in the challenge** sets `severity_avg = 5` (the maximum), for backward compatibility with `csi-v0`, which had no severity field.

The second convention is a known perverse incentive — an all-omitted challenge scores its assumption component higher than an identical challenge with honestly declared mid severities. It is documented as an open weakness, not silently accepted (it is a disclosed weakness, not a hidden one). Production pipelines SHOULD always declare severities.

`severity_avg ∈ [1, 5]`. The modulation factor applied to the assumption component is `severity_avg / 5 ∈ [0.2, 1.0]`.

A.6 — THE FIVE COMPONENTS

With core weights `w = (w_a, w_F, w_B, w_S, w_D) = (25, 25, 20, 10, 20)`, `Σ w = 100`:

COMPONENT	FORMULA	MAX
assumptions	`c_a = w_a · norm(n_a) · (severity_avg / 5)`	25
falsifications	`c_F = w_F · norm(n_F)`	25
burden questions	`c_B = w_B · norm(n_B)`	20
sources	`c_S = w_S · norm(n_S)`	10
divergence	`c_D = w_D · d`, with `d ∈ [0, 1]` (A.9)	20

Each component is independent; there are no cross-terms. Additivity is deliberate: the score is decomposable, so any Δ-CSI can be attributed component by component.

A conforming implementation MAY override the weights via configuration (a "Pack"), subject to the sum constraint (A.10.4). It MUST NOT change the functional form of any component, nor the cap, nor the bands.

A.7 — AGGREGATION AND ROUNDING

`Δ-CSI = round₂( cₐ + c_F + c_B + c_S + c_D )`

where `round₂` is round-half-to-even to two decimal places for the score. The divergence scalar `d` is carried to four decimal places in the record. `Δ-CSI ∈ [0, 100]`; the maximum 100 is attained only when every component is simultaneously maximal (all counts ≥ 5, all severities = 5, `d = 1`).

A.8 — INTERPRETATION BANDS

Three fixed bands, identical across every domain and configuration:

BAND	INTERVAL
marginal	`[0, 33)`
moderate	`[33, 66)`
substantial	`[66, 100]`

The bands are fixed by specification and MUST NOT be moved by any Pack. The rationale: if installations could move the thresholds, the word "moderate" would lose its shared meaning and the same number would tell different stories to different readers. The canonical interpretation strings are of the form: *"The contradictor applied {marginal | moderate | substantial} pressure to the decision's assumptions."* — a statement about challenge pressure, never about the decision's merit or about a change in the decision-maker's judgment.

A.9 — DIVERGENCE

The divergence scalar `d ∈ [0, 1]` measures how far the counter-intuitive scenario `κ` departs from the thesis `τ`. Two methods are defined; both return `[0, 1]` but measure **different properties** and are therefore not mutually comparable (A.12).

A.9.1 — Heuristic (Jaccard, lexical).

Let `T(x)` be the token set of text `x` (A.2). Then:

`d_jaccard = 1 − |T(τ) ∩ T(κ)| / |T(τ) ∪ T(κ)|`

Deterministic, offline, dependency-free. It is the guaranteed fallback in any configuration. It measures **lexical** distance: texts that agree in meaning but differ in wording score as divergent.

A.9.2 — Semantic (cosine).

Given a semantic embedding provider `embed(·)` returning vectors, with cosine similarity `cos`:

`d_cosine = (1 - cos( embed(τ), embed(κ) )) / 2`

mapping cosine from `[-1, 1]` to `[0, 1]`. It measures **semantic** distance and is more faithful to the construct, at the cost of a dependency on an external encoder. Because the encoder varies by deployment, cosine scores from different encoders are not directly comparable (A.12).

A.9.3 — Method selection and the recorded method.

Selection reads an environment setting `BRAIN_DIVERGENCE ∈ {auto, cosine, heuristic}` (default `auto`) and the availability of a semantic embedding provider:

SETTING	SEMANTIC PROVIDER	METHOD USED
`auto`	available	cosine
`auto`	absent	Jaccard
`cosine`	available	cosine
`cosine`	absent	Jaccard (warning logged)
`heuristic`	any	Jaccard

Fail-safe. If the encoder raises during embedding, the cosine strategy falls back silently to Jaccard; the pipeline never crashes on an encoder error. The record's `divergence_method` field MUST reflect the method **actually executed** on that call (`heuristic` on fallback), not the nominally selected one, so calibration cohorts stay pure.

A.10 — INVARIANTS (A CONFORMING IMPLEMENTATION MUST NOT BREAK THESE)

A.10.1 — Fixed output schema. Every analysis MUST emit the seven contradictor sections — implicit assumptions · counter-intuitive scenario · falsification tests · burden-of-proof questions · calibrated confidence · cited sources · Δ-CSI — plus the service fields (`schema_version`, `decision`, `thesis`, `meta`). The contract is `schema_version = "1.0"` with `additionalProperties: false` on every object. A Pack MAY change voice, tone, and lexicon; it MUST NOT remove, rename, or make optional any section.

A.10.2 — Contradiction floor (fail-closed). A challenge is admissible only if it has ≥ 1 assumption, ≥ 1 falsification test, ≥ 1 burden question, and a non-empty counter-scenario narrative. If the floor is unmet after the pass that produces the challenge, the implementation MUST perform exactly one sharpened retry; if still unmet, it MUST raise a controlled error and emit no score. A reassuring score is never emitted in place of an empty challenge.

A.10.3 — Closed provenance enum. Every source MUST declare a provenance drawn from the closed set `{web_search, internal_ledger, internal_doc, osint_module, user_provided}`. With no retrieval available, the synthesis MUST return `sources: []`; sources are never fabricated.

A.10.4 — Sum-to-100 weights. Any weight override MUST satisfy `Σ w = 100`, enforced programmatically. This forbids covert inflation of the maximum, which stays 100.

A.10.5 — Honest calibration gate. Below `MIN_SAMPLE = 5` outcome-bearing records in a cohort, calibration output MUST be flagged descriptive-only (`sufficient = false`) and no validity conclusion is permitted (A.12).

A.11 — REPRODUCIBLE WORKED EXAMPLE (GOLDEN)

Inputs: $n_a = 3$ with no declared severities (→ `severity_avg = 5.0`); `n_F = 2`; `n_B = 2`; `n_S = 1`; `d = 0.9615384615384616` (heuristic Jaccard).

COMPONENT	N / VALUE	ARITHMETIC	CONTRIBUTION
assumptions	3, `severity_avg` 5.0	`25 · (3/5) · (5/5)`	15.00
falsifications	2	`25 · (2/5)`	10.00
burden questions	2	`20 · (2/5)`	8.00
sources	1	`10 · (1/5)`	2.00
divergence	0.9615	`20 · 0.9615…`	19.23
Δ-CSI			**54.23**

`divergence = 0.9615384615384616 score = 54.23 band = moderate`

These numbers are produced by the reference scorer `score_delta_csi` (`csi-v1`) and reproduce deterministically: given the same components, the scorer always returns the same number. The narrative text of such a challenge is not deterministic (it is produced by a language model); only the quantitative part is reproducible by construction.

A.12 — COMPARABILITY AND CALIBRATION

Cohort key. Δ-CSI records are comparable only within a cohort defined by the triple:

`cohort = (version, divergence_method, weights)`

Scores produced with different computation versions (`csi-v0` vs `csi-v1`), different divergence methods (Jaccard vs cosine), or different weight distributions are **not** comparable and MUST NOT be pooled. A Δ-CSI of 54 under Jaccard and a Δ-CSI of 54 under cosine do not assert the same thing about challenge intensity; the numerical identity is accidental.

Calibration. Calibration computes descriptive statistics only, per cohort. Below `MIN_SAMPLE = 5` outcome-bearing records (outcomes ≠ "unknown"), `sufficient = false` and no validity claim is licensed. Above the threshold, a `hit_rate` MAY be produced:

`hit_rate = (changed decisions with correct outcome) / (changed decisions with definitive outcome)`

`hit_rate` may legitimately be `None` even above threshold (zero denominator when no decision was ever changed). It is per-tenant, descriptive, local evidence; it is not causal, and does not generalize across tenants without further analysis. This engine-level gate (≥ 5) is a product safeguard; it is independent of, and far weaker than, any confirmatory statistical power floor used in an empirical validation study.

A.13 — VERSIONING OF THIS SPECIFICATION

`spec-1.0` binds `csi-v1`. A change to any formula, weight default, band, cap, invariant, cohort key, or the `MIN_SAMPLE` constant requires a specification version bump and a new computation version tag. Records always carry the computation version (`csi_delta.version`), so historical scores remain interpretable under the version that produced them.

REFERENCES

References

1 Bainbridge, L. (1983). Ironies of automation. *Automatica*, 19(6), 775–779. https://doi.org/10.1016/0005-1098(83)90046-8

2 Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1), 1–3. https://doi.org/10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2

3 Cai, A., Arawjo, I., & Glassman, E. L. (2024). *Antagonistic AI* (arXiv:2402.07350) [Preprint]. arXiv. https://arxiv.org/abs/2402.07350

4 Canepa, F. S. (2026a). *Cognitive Sovereignty: Protecting the Quality of Leadership Decisions in the Age of Artificial Intelligence* [Preprint]. OSF. https://osf.io/6n8tg

5 Canepa, F. S. (2026b). *Calibrated Dissent: A Verifiable Sealing-and-Resolution Protocol for Auditing Adversarial AI Forecasts, with a Public M&A Pilot (N = 10)* [Preprint]. OSF. https://doi.org/10.17605/OSF.IO/5QC8T

6 Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, 52(4), 281–302. https://doi.org/10.1037/h0040957

7 De Luca, V., Giamminola, G., & Pollicino, O. (2026). *Il Riarmo Cognitivo: La sicurezza cognitiva come struttura dual use* [Cognitive Rearmament: Cognitive security as a dual-use structure] [Synthesis report, June 2026]. Rome. *(The organizational Cognitive Sovereignty Index is introduced in this work; publisher/venue to be finalized for the record of publication.)*

8 Goodhart, C. A. E. (1975). Problems of monetary management: The U.K. experience. In *Papers in Monetary Economics* (Vol. 1). Reserve Bank of Australia. [Reprinted in C. A. E. Goodhart (1984), *Monetary Theory and Practice: The U.K. Experience*, pp. 91–121, Macmillan.]

9 Howard, R. A. (1968). The foundations of decision analysis. *IEEE Transactions on Systems Science and Cybernetics*, 4(3), 211–219. https://doi.org/10.1109/TSSC.1968.300115

10 Jaccard, P. (1912). The distribution of the flora in the alpine zone. *New Phytologist*, 11(2), 37–50. https://doi.org/10.1111/j.1469-8137.1912.tb05611.x

11 Kahneman, D. (2011). *Thinking, Fast and Slow*. Farrar, Straus and Giroux.

12 Kleinberg, J., & Raghavan, M. (2021). Algorithmic monoculture and social welfare. *Proceedings of the National Academy of Sciences*, 118(22), e2018340118. https://doi.org/10.1073/pnas.2018340118

13 Mayo, D. G. (1996). *Error and the Growth of Experimental Knowledge*. University of Chicago Press.

14 Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, 50(9), 741–749. https://doi.org/10.1037/0003-066X.50.9.741

15 Murphy, A. H. (1973). A new vector partition of the probability score. *Journal of Applied Meteorology*, 12(4), 595–600. https://doi.org/10.1175/1520-0450(1973)012<0595:ANVPOT>2.0.CO;2

16 Parasuraman, R., & Manzey, D. H. (2010). Complacency and bias in human use of automation: An attentional integration. *Human Factors*, 52(3), 381–410. https://doi.org/10.1177/0018720810376055

17 Popper, K. R. (1959). *The Logic of Scientific Discovery*. Hutchinson. [Original work published as *Logik der Forschung*, 1935.]

18 Salton, G., Wong, A., & Yang, C. S. (1975). A vector space model for automatic indexing. *Communications of the ACM*, 18(11), 613–620. https://doi.org/10.1145/361219.361220

19 Schwenk, C. R. (1990). Effects of devil's advocacy and dialectical inquiry on decision making: A meta-analysis. *Organizational Behavior and Human Decision Processes*, 47(1), 161–176. https://doi.org/10.1016/0749-5978(90)90051-A

20 Sharma, M., Tong, M., Korbak, T., Duvenaud, D., Askell, A., Bowman, S. R., … Perez, E. (2023). *Towards Understanding Sycophancy in Language Models* (arXiv:2310.13548) [Preprint]. arXiv. Published at ICLR 2024. https://arxiv.org/abs/2310.13548

21 Steenbergen, M. R., Bächtiger, A., Spörndli, M., & Steiner, J. (2003). Measuring political deliberation: A Discourse Quality Index. *Comparative European Politics*, 1(1), 21–48. https://doi.org/10.1057/palgrave.cep.6110002

22 Strathern, M. (1997). 'Improving ratings': Audit in the British University system. *European Review*, 5(3), 305–321. https://doi.org/10.1002/(SICI)1234-981X(199707)5:3<305::AID-EURO184>3.0.CO;2-4

23 Tetlock, P. E., & Gardner, D. (2015). *Superforecasting: The Art and Science of Prediction*. Crown.

24 Wachsmuth, H., Naderi, N., Hou, Y., Bilu, Y., Prabhakaran, V., Alberdingk Thijm, T., Hirst, G., & Stein, B. (2017). Computational argumentation quality assessment in natural language. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2017), Volume 1: Long Papers* (pp. 176–187). Association for Computational Linguistics. https://aclanthology.org/E17-1017/

25 Wu, H., Li, Z., & Li, L. (2025). *Can LLM Agents Really Debate? A Controlled Study of Multi-Agent Debate in Logical Reasoning* (arXiv:2511.07784) [Preprint]. arXiv. https://arxiv.org/abs/2511.07784

26 Wynn, A., Satija, H., & Hadfield, G. (2025). *Talk Isn't Always Cheap: Understanding Failure Modes in Multi-Agent Debate* (arXiv:2509.05396) [Preprint]. arXiv. https://arxiv.org/abs/2509.05396

27 Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E. P., Zhang, H., Gonzalez, J. E., & Stoica, I. (2023). *Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena* (arXiv:2306.05685) [Preprint]. arXiv. Published in *Advances in Neural Information Processing Systems 36* (NeurIPS 2023), Datasets and Benchmarks Track. https://arxiv.org/abs/2306.05685