PAPER DI COSTRUTTO E METODO DI MISURA

Misurare la contestazione, non la decisione

Un indice deterministico dell'intensità della sfida in un contraddittore cognitivo AI — il Δ-CSI (Cognitive Sovereignty Index, delta)

di

Francesco Saverio Canepa

Ricercatore indipendente, Roma

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

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

INDICE

Indice

PAPER DI COSTRUTTO E METODO DI MISURA

Sintesi

Chi decide con l'AI come assistente intensivo si scontra con un paradosso ormai documentato: più analisi non porta, di per sé, a un giudizio migliore. Gli assistenti ottimizzati per compiacere l'utente tendono a confermare invece che a contestare; così gonfiano la sicurezza di chi decide e uniformano il ragionamento tra le organizzazioni che li condividono. Esiste un'alternativa, ed è strutturalmente avversariale: un *contraddittore cognitivo*, che ha il compito di sfidare una decisione, non di avallarla. Ma quell'alternativa solleva subito una domanda di misura. Con quanta intensità una decisione è stata contestata — e come si risponde in un modo che sia riproducibile nel punteggio, confrontabile nel tempo e onesto su ciò che il numero dice e su ciò che tace?

Questo paper definisce quella grandezza. Introduce il **Δ-CSI (Cognitive Sovereignty Index, delta)** come costrutto — l'*intensità della sfida strutturata* applicata a una singola decisione in una singola sessione di analisi — e ne fissa la definizione operativa.

È un indice deterministico in [0, 100], calcolato da cinque componenti pesati. Sono il conteggio delle assunzioni implicite modulato dalla loro gravità media, il conteggio dei test di falsificazione, quello delle domande d'onere della prova, quello delle fonti citate, e la divergenza tra tesi e contro-scenario. I conteggi hanno un tetto, che impedisce di gonfiare il punteggio con la sola verbosità. I pesi sono vincolati a sommare a cento, le fasce di interpretazione sono fisse e lo schema di versione è esplicito. Restano però due limiti, dichiarati apertamente: il canale della gravità auto-dichiarata è aggirabile per costruzione (§7), e i componenti di conteggio registrano la quantità, non la qualità degli argomenti.

Il contributo è triplice. (i) Il costrutto, delimitato da quattro negazioni: l'intensità della sfida *non* è qualità della decisione, *non* è probabilità di successo, *non* è un giudizio sul decisore, *non* è il fatto che la decisione sia poi cambiata. (ii) La specifica, con la sua garanzia di riproducibilità, le difese fail-closed contro l'output compiacente e un registro esplicito delle debolezze. (iii) La disciplina di confrontabilità e calibrazione: coorti indicizzate per versione del calcolo, metodo di divergenza e pesi, con statistiche tenute puramente descrittive sotto i cinque esiti registrati.

Il paper non avanza alcuna pretesa di validità empirica o di efficacia. Non afferma che un Δ-CSI alto corrisponda a decisioni migliori — né, all'opposto, a decisioni più deboli — e non afferma che l'indice predica gli esiti: dimostra soltanto la definizione del costrutto e il metodo di misura.

La validazione empirica è dichiarata come programma differito, strutturato come una roadmap pre-registrata di accumulo dell'evidenza. Le due tappe sono la calibrazione del ledger per-tenant e il pilot pubblico e sigillato di M&A del companion *Calibrated Dissent*, dove il Δ-CSI entra soltanto come provenienza — senza concorrere al punteggio — e come

covariata esplorativa. Trattiamo questa onestà come un requisito di progettazione, non come una clausola di cautela: un indice dell'intensità della sfida che sopravvalutasse la propria validazione contraddirebbe la postura stessa che esiste per misurare.

Parole chiave: qualità della decisione; contraddittore cognitivo; AI avversariale; validità di costrutto; intensità della sfida; calibrazione; governance dell'AI; misura.

SEZIONE 1

Introduzione

Chi guida un'organizzazione dispone oggi di un'assistenza analitica senza precedenti. I sistemi AI sintetizzano evidenze, costruiscono scenari, restituiscono argomenti più in fretta e meglio rifiniti di qualsiasi ufficio studi.

Da qui si deduce, quasi sempre senza verificarlo, che decisioni così assistite siano decisioni migliori. C'è motivo di dubitarne.

Il modello del mondo su cui poggia una decisione consequenziale viene raramente messo alla prova prima di diventare azione. Più spesso viene confermato in silenzio. Cercare evidenze a favore dell'ipotesi che già si preferisce è tra i risultati più solidi dello studio del giudizio in condizioni di incertezza (Kahneman, 2011). E l'assistente più gradevole da consultare è esattamente quello che quelle evidenze le fornisce a richiesta.

Il meccanismo è documentato, non congetturato. I modelli linguistici messi a punto con feedback umano mostrano *sycophancy*: una tendenza misurabile ad allinearsi alla posizione dell'utente, fino a rivedere risposte corrette verso l'errore dell'utente quando incalzati (Sharma et al., 2023).

Il comportamento aggrava due risultati più antichi dello studio dell'interazione uomo-automazione. L'*over-reliance* è la tendenza ad accettare una raccomandazione automatica senza il vaglio che si applicherebbe a una proposta umana (Parasuraman & Manzey, 2010). Il *deskilling* è la perdita di competenza che segue quando se ne delega progressivamente l'esercizio a un sistema che la svolge (Bainbridge, 1983).

Un quarto effetto opera tra le organizzazioni anziché al loro interno, perché tutte consultano lo stesso ristretto numero di modelli: la *monocultura algoritmica*. Strumenti correlati inducono giudizi correlati, e gli errori del sistema condiviso smettono di essere indipendenti tra le istituzioni che vi fanno affidamento (Kleinberg & Raghavan, 2021).

Il paradosso, formulato con precisione: l'analisi abbondante può aggravare il fallimento che sembra rimediare. Più la conferma è fluente, meno la decisione è messa alla prova — e più le sue premesse non verificate sono le stesse di chiunque usi lo stesso strumento.

Il companion *Sovranità cognitiva* (Canepa, 2026a) sviluppa il costrutto che nomina ciò che si erode in queste condizioni: la capacità collettiva di tenere il giudizio indipendente, calibrato e falsificabile. E identifica il meccanismo che lo proteggerebbe — un sistema la cui funzione è contestare una decisione, non assistere il ragionamento di chi l'ha presa.

Sulla scia di quel lavoro chiamiamo il meccanismo **contraddittore cognitivo**, e usiamo il termine per tutto il testo. Le etichette più tenui non servono: «critico», «consulente», «avvocato del diavolo» ammettono per connotazione proprio la postura opzionale e consultiva che il meccanismo è progettato per escludere. L'output di un contraddittore *è* la sfida: contestare è la sua funzione-obiettivo, non un sottoprodotto occasionale di un'analisi orientata alla sintesi.

Un meccanismo definito dall'intensità della sfida che applica solleva una domanda ovvia, ed è una domanda di misura. Se la contestazione è il punto, si deve poter dire *quanta* ce ne sia stata. Altrimenti l'affermazione che una data analisi fosse avversariale diventa infalsificabile, e la distinzione tra una sfida sostanziale e la sua messinscena si dissolve.

La grandezza in questione non è la qualità della decisione: inosservabile nel momento in cui si decide, e confusa dalla fortuna in seguito. Non è nemmeno l'accuratezza previsiva, che è una proprietà distinta, misurata con altri mezzi (Brier, 1950) e sottoposta ad audit nel protocollo companion (Canepa, 2026b). È l'*intensità della sfida stessa*: con quanta forza, e lungo quante dimensioni indipendenti, il contraddittore ha premuto sulle assunzioni della decisione in esame.

Questo paper definisce quella grandezza e ne specifica la misura. Chiamiamo l'indice **Δ-CSI — Cognitive Sovereignty Index, delta**. La scelta del nome è deliberata, ed è la prima cosa che un lettore attento interrogherà (§3.4).

Il contributo è triplice. Primo, il *costrutto*: l'intensità della sfida, delimitata da quattro nozioni adiacenti con cui è abitualmente confusa, e non senza conseguenze (§3). Secondo, la *definizione operativa*: un indice deterministico in [0, 100], con i suoi cinque componenti, le sue regole anti-gaming e di confrontabilità, la sua garanzia di riproducibilità e le sue difese fail-closed, fissate in una specifica versionata (§4–§7). Terzo, la *disciplina di validazione*: un resoconto preciso di ciò che l'indice stabilisce e di ciò che non stabilisce, con una roadmap pre-registrata lungo cui la validità empirica potrebbe essere accumulata — mai assunta (§6, §8).

La postura del paper sulla propria validazione non è modestia. È un impegno metodologico, ed è portante.

Un metodo di misura presentato in prosa fluente e sicura è particolarmente esposto al resoconto selettivo e alla sopravvalutazione. E un progetto la cui premessa è che proprio della conferma fluente occorra diffidare non può esentare dal sospetto il proprio resoconto.

Prendiamo la lezione alla lettera. Ogni affermazione fattuale qui sotto è ancorata alla specifica, all'implementazione di riferimento o a una fonte citata. Ogni proprietà non ancora dimostrata è marcata come tale. Il confine tra ciò che è definito, ciò che è ingegnerizzato e ciò che resta congettura empirica è tracciato esplicitamente e tenuto visibile.

Un costrutto il cui scopo è misurare l'intensità dell'auto-contestazione non può essere introdotto da un documento che si sottrae alla stessa contestazione. La Sezione 7 rivolge perciò il metodo sul proprio indice, e registra senza mitigazione le debolezze che sopravvivono.

Il resto del paper procede come segue. La Sezione 2 posiziona l'indice rispetto alle letterature da cui attinge e alle misure con cui non va confuso. La Sezione 3 definisce il costrutto e le sue quattro negazioni. La Sezione 4 dà la definizione operativa. La Sezione 5 colloca lo scorer nella pipeline che ne produce gli input, ed enuncia la separazione provenienza/scoring che mantiene interpretabile l'indice.

La Sezione 6 descrive la persistenza e la disciplina di calibrazione: il senso, preciso e limitato, in cui l'indice è «calibrato». La Sezione 7 è l'auto-contraddittorio, cioè le debolezze

dichiarate. La Sezione 8 enuncia i limiti e la roadmap di validazione pre-registrata. La Sezione 9 fissa i termini su cui la specifica è offerta come standard aperto. La Sezione 10 conclude.

SEZIONE 2

Related work e posizionamento

Il Δ-CSI si colloca all'intersezione di quattro letterature e va distinto da una quinta famiglia di misure con cui condivide il lessico ma non l'oggetto. Le prendiamo una per una, perché il contributo del paper è leggibile solo rispetto a ciò che *non* è.

Sfida strutturata nelle decisioni. Che la qualità di una decisione consequenziale dipenda dall'architettura del processo che l'ha prodotta è un risultato antico e ben sostenuto. Non dipende dalla sola competenza dei partecipanti, né dalla quantità di informazioni disponibili.

La meta-analisi di Schwenk (1990) su devil's advocacy e inquiry dialettico ha fornito evidenza che il disaccordo strutturato può migliorare la qualità della decisione rispetto alla ricerca del consenso, con effetti contingenti al modo in cui la procedura è implementata. E ha mostrato, cosa cruciale, che l'effetto è una proprietà della *procedura* piuttosto che degli individui che la mettono in atto.

Il contraddittore cognitivo eredita direttamente questa lezione: il mandato di sfidare è codificato nell'architettura del sistema, non delegato alla vigilanza o al carattere di un avvocato umano che può stancarsi, cedere o essere messo in minoranza. Il Δ-CSI è lo strumento che rende osservabile — e quindi verificabile — l'intensità di quella sfida codificata, sessione per sessione.

AI avversariale e antagonistica. Una linea di lavoro recente sostiene che il default progettuale quasi universale — un'AI accomodante e compiacente — non è né inevitabile né sempre desiderabile, ed esplora l'interazione deliberatamente antagonistica come spazio di progetto (Cai, Arawjo, & Glassman, 2024). Il contraddittore è un'istanza di quella postura orientata alla governance: ristretta a una sola funzione, contestare una decisione, e vincolata a un contratto di output fisso.

Ma è la stessa letteratura a dare la ragione per cui la sfida va *misurata*, non solo *invocata*. Studi controllati sul dibattito multi-agente mostrano che le cornici avversariali non producono in modo attendibile un disaccordo genuino. Gli agenti spesso convergono, ripetono un prior condiviso, o mettono in scena l'opposizione senza cambiare la sostanza del loro ragionamento (Wynn, Satija, & Hadfield, 2025; Wu, Li, & Li, 2025).

Un indice dell'intensità della sfida è, in parte, una difesa contro questo guasto. È però una difesa solo parziale, e il paper lo dice apertamente.

Un'opposizione simulata che sia anche *sottile* — poche assunzioni, gravità dichiarata bassa, un contro-scenario vicino alla tesi — ottiene in effetti un punteggio basso. Un'altra però sfugge all'indice: una messinscena che riproduce i marcatori strutturali di una sfida senza la sua sostanza, con molte assunzioni auto-valutate come gravi, test verbosi ma banali, un contro-scenario lontano nelle parole ma non davvero opposto.

La costruzione attuale non distingue quella messinscena da una sfida genuina. Quattro componenti su cinque registrano la struttura — tre sono conteggi puri, più la distanza strutturale della divergenza — invece di valutare la qualità, e l'unico segnale di qualità resta la gravità delle assunzioni (§3.3, §7.2).

L'indice alza il costo dell'opposizione simulata più grossolana; non intercetta quella sofisticata. Quel divario è una delle debolezze dichiarate al §7.2, non una proprietà che l'indice pretenda di avere.

Calibrazione e forecasting. La letteratura del forecasting offre sia una disciplina che il Δ-CSI adotta, sia una misura con cui non va confuso.

La calibrazione, nel senso di Tetlock e Gardner (2015) e della tradizione di scoring che li precede (Brier, 1950; Murphy, 1973), è la corrispondenza tra il grado di confidenza dichiarato e la frequenza empirica degli esiti. È una proprietà delle *previsioni*, stabilita confrontando probabilità ed eventi realizzati. Il Δ-CSI non è una previsione e non ha alcun Brier score: non avanza alcuna pretesa probabilistica su alcun esito.

Dalla tradizione del forecasting l'indice prende in prestito l'*umiltà epistemica*, cioè il rifiuto di affermare validità predittiva prima dei dati di esito. E ne usa il macchinario solo a valle, nella disciplina di calibrazione del §6, dove i punteggi Δ-CSI accumulati sono confrontati con gli esiti decisionali registrati per ciascun tenant.

Il protocollo companion (Canepa, 2026b) è la sede in cui le *previsioni* del contraddittore sono sottoposte ad audit rispetto agli esiti, secondo un piano pre-registrato. Lì il Δ-CSI entra come provenienza — registrata, ma non conteggiata nel punteggio — mai come grandezza a punteggio. La separazione è deliberata e vale in tutto il corpus: questo paper misura la sfida, il companion verifica la previsione.

Validità di costrutto. Poiché il Δ-CSI è una misura proposta di una proprietà inosservabile, appartiene alla tradizione della validazione di costrutto nelle scienze della misura (Cronbach & Meehl, 1955; Messick, 1995). Quella tradizione fornisce lo standard di adeguatezza del paper e la sua cautela principale.

Cronbach e Meehl richiedono che una misura di costrutto sia inserita in una *rete nomologica* di relazioni teoriche, contro cui il suo comportamento possa essere messo alla prova. È un requisito che il corpus, al presente, può solo *abbozzare*, non soddisfare. Il paper di costrutto (Canepa, 2026a) colloca l'intensità della sfida tra le dimensioni della sovranità cognitiva; questo paper la operativizza; il protocollo companion (Canepa, 2026b) specifica i test empirici.

Ma le relazioni che questo paper può enunciare oggi sono, in onestà, per lo più *discriminanti* — ciò a cui l'indice *non* è legato (§3.2) — e definitorie anziché dimostrate. I test *convergenti* che confermerebbero il costrutto sono esposti nella roadmap (§8.2) e non eseguiti qui. Sarebbero, per esempio, l'accordo tra il Δ-CSI e valutazioni indipendenti di esperti sulle stesse sfide, o una separazione a gruppi noti di sfide genuine da sfide deliberatamente inscenate. Enunciamo la rete come un'aspirazione con un programma confermativo differito, non come una conquista.

Messick (1995) insiste sul fatto che la validità sia una proprietà delle *inferenze tratte dai punteggi*, non dello strumento in quanto tale. È ciò che detta la restrizione più importante del paper: non validiamo qui alcuna inferenza. Definiamo il punteggio e la sua aritmetica; quali inferenze quel punteggio autorizzerà è una questione empirica, riservata alla roadmap del §8.

Misurare la qualità dell'argomentazione e della deliberazione. Il Δ-CSI conta e pesa mosse dialettiche — assunzioni, test di falsificazione, domande d'onere della prova — e ha perciò un'ascendenza in due tradizioni di misura che il paper deve riconoscere, perché sono dove un referee guarderà per primo.

Nella ricerca sulla deliberazione, il Discourse Quality Index (Steenbergen, Bächtiger, Spörndli, & Steiner, 2003) valuta la qualità della deliberazione politica lungo dimensioni codificate di giustificazione e rispetto. Nella linguistica computazionale, l'argument mining e la valutazione della qualità argomentativa (Wachsmuth et al., 2017) costruiscono tassonomie per graduare la fondatezza e la ragionevolezza degli argomenti in linguaggio naturale.

Il Δ-CSI differisce da entrambi nell'oggetto, e la differenza è deliberata anziché una svista. Quegli strumenti graduano la *qualità* dell'argomentazione, tipicamente tramite codifica umana o di modelli addestrati. Il Δ-CSI misura l'*intensità* della contestazione tramite conteggi in prevalenza strutturali, precisamente affinché la sua parte quantitativa sia riproducibile senza uno schema di codifica (§3.3, §4). Guadagna riproducibilità al costo della sensibilità alla qualità — uno scambio che la letteratura sulla qualità argomentativa fa in direzione opposta — e il §7.2 registra il limite che ne risulta anziché mascherarlo.

Adiacente, inoltre, è la tradizione dell'analisi delle decisioni che colloca la qualità di una decisione nel suo *processo* piuttosto che nel suo esito (Howard, 1968). Il Δ-CSI strumenta una faccetta di quel processo — l'intensità della sfida applicata — senza pretendere di misurare la qualità della decisione stessa (§3.2).

Ciò che il Δ-CSI non è: il CSI organizzativo. Una misura dal nome quasi identico compare nella letteratura circostante sulla sicurezza cognitiva, e la collisione è abbastanza consequenziale da richiedere un paragrafo anziché una nota.

Il *CSI organizzativo* è stato introdotto nel programma «Riarmo Cognitivo» (De Luca, Giamminola, & Pollicino, 2026). Indicizza la postura cognitiva di un'intera organizzazione lungo dimensioni strutturali: la sua dipendenza dai sistemi algoritmici, il grado in cui quei sistemi inquadrano le decisioni prima che siano prese consapevolmente, e la capacità critica residua dell'organizzazione. È un costrutto a livello di framework, valutato su un'istituzione.

Il Δ-CSI è un oggetto interamente diverso. È deterministico dove il CSI organizzativo è framework-based, opera su una singola decisione dove il CSI organizzativo opera su un'organizzazione, e misura l'intensità di una singola sfida dove il CSI organizzativo misura una postura stabile. I due indici condividono tre lettere e una genealogia di preoccupazione; non condividono un livello di analisi, un metodo di costruzione, né un oggetto di misura.

Un ponte concettuale tra il livello-sessione e il livello-organizzazione è plausibile: decisioni abitualmente sottoposte a contestazione intensa potrebbero, nel tempo, rafforzare la postura cognitiva di un'organizzazione. Ma nessun ponte simile è formalizzato, nessuna mappatura calibrata tra i due indici esiste, e nulla in questo paper va letto come se ne esistesse una.

Dove questo paper scrive «CSI» senza qualificazione si riferisce all'indice organizzativo di De Luca et al. (2026); l'indice a livello-sessione è sempre scritto «Δ-CSI».

Il posizionamento può riassumersi in una singola disciplina. Il Δ-CSI prende in prestito l'intuizione *procedurale* della letteratura sulla sfida strutturata. Istanzia la postura *avversariale*, difendendosi dal suo noto fallimento nel produrre disaccordo reale. Adotta l'*umiltà* della tradizione della calibrazione senza pretenderne il macchinario predittivo. E si sottopone allo standard della *validità di costrutto*, differendo esplicitamente la validazione di qualsiasi inferenza. Ciò che resta è definire il costrutto con precisione sufficiente perché questi impegni abbiano mordente.

SEZIONE 3

Il costrutto: intensità della sfida

3.1 DEFINIZIONE

Il Δ-CSI misura l'**intensità della sfida strutturata** che un contraddittore cognitivo applica alle assunzioni di una singola decisione, in una singola sessione di analisi.

«Intensità» è qui un composto di estensione e forza. L'estensione: quanti punti indipendenti di contestazione la sfida ha sollevato — assunzioni fatte emergere, test di falsificazione proposti, oneri della prova assegnati, fonti addotte. La forza: quanto il contro-scenario si è allontanato dalla tesi che opponeva, ponderato per quanto gravi fossero le assunzioni emerse.

È una proprietà della *sfida*, prodotta in una sessione, su una decisione. Non è una proprietà della decisione, del suo futuro, o della persona che l'ha presa.

È un oggetto deliberatamente modesto: l'indice non aggiudica la disputa che misura, la misura e basta.

Un'analogia fissa il registro. Un pugile che prepara un incontro non chiede allo sparring partner di prevedere il risultato: chiede di essere messo alla prova. A fine round può dire quanto è stato duro — quanti colpi ha dovuto parare, quante volte è finito in difficoltà. Quel giudizio misura l'intensità della sessione e non pronostica nulla sull'incontro.

Il Δ-CSI è quel giudizio reso in numero: quanto è stato intenso il confronto tra una decisione e il suo contraddittore. Non dice se la decisione vincerà.

3.2 LE QUATTRO NEGAZIONI

Il costrutto è definito tanto da ciò che esclude quanto da ciò che afferma. Quattro esclusioni sono costitutive. Ciascuna corrisponde a un errore di categoria che il nome dell'indice e la sua scala numerica invitano attivamente a commettere.

Non la qualità della decisione. Un Δ-CSI alto non significa che la decisione sia sbagliata; uno basso non significa che sia giusta. L'indice dà un punteggio alla sfida, non alla scelta. Una decisione può uscire intatta da un contraddittore intenso — punteggio alto, decisione solida — oppure crollare sotto uno mite. Il numero non distingue i due casi, perché non riguarda affatto la decisione.

Non la probabilità di successo. Nessun componente del calcolo stima come andrà a finire la decisione. L'indice è silente sugli esiti per costruzione: legge la struttura della sfida, e la struttura di una sfida non contiene alcuna previsione. L'accuratezza previsiva è una proprietà distinta, con strumenti propri (Brier, 1950), ed è sottoposta ad audit altrove nel corpus (Canepa, 2026b).

Non un giudizio sul decisore. L'indice non è un voto per la persona o il gruppo che ha formulato la decisione, né una valutazione del lavoro fatto prima dell'analisi. Una decisione mal formulata può ottenere un punteggio alto proprio perché ha offerto al contraddittore appigli abbondanti. Una ben formulata può ottenerne uno basso perché restava poco da contestare. Il numero descrive la contesa, non il contendente.

Non se la decisione sia cambiata. Cosa più consequenziale: il Δ-CSI non misura se il contraddittore abbia effettivamente spostato il giudizio del decisore. Quell'effetto — se l'analisi abbia cambiato la decisione prima che fosse attuata — è un fatto separato e osservabile. Il ledger lo registra a parte, come campo booleano (`decision_changed`), e non lo fonde mai nell'indice.

Intensità della sfida e cambiamento del giudizio sono cose distinte. Una sfida intensa può lasciare la decisione invariata: il decisore ha considerato le obiezioni e ha tenuto la posizione. Una mite può cambiarla: è bastata una sola domanda ben piazzata. Confonderle renderebbe l'indice una misura della propria persuasività, che non è ciò che misura.

Queste non sono cautele appese a una misura che, senza di esse, pretenderebbe di dire quelle cose: sono il confine del costrutto. Un errore di categoria non si corregge con più precisione. Chiedere al Δ-CSI se una decisione sia solida è come chiedere a un termometro se il tempo sia gradevole, e nessuna soglia sul termometro risponde alla domanda. Tenere l'intensità della sfida separata dalla qualità della decisione, dalla previsione, dal merito e dalla persuasione è la condizione perché l'indice sia utile anziché fuorviante.

3.3 UN PROXY, E DI QUALE TIPO

Il Δ-CSI è un proxy: una quantificazione indiretta di una proprietà inosservabile attraverso segnali osservabili. È cosa ordinaria nelle scienze della misura, dove la maggior parte dei costrutti di interesse è latente e si raggiunge per proxy. Ma il *tipo* di proxy conta. La validità di costrutto (Cronbach & Meehl, 1955) ci impone di enunciare due conseguenze.

Primo: l'indice misura l'*intensità della sfida prodotta*, non la *qualità degli argomenti* al suo interno.

Tre dei suoi cinque componenti sono conteggi (§4). Un test di falsificazione contribuisce al punteggio sia che sia incisivo, sia che sia banale. Lo scorer conta gli elementi che la pipeline etichetta come test di falsificazione, ma non verifica che ciascuno sia un test genuino e con criterio anziché una semplice esortazione — un limite ereditato dall'impianto a soli conteggi (§7.2).

Il costrutto dichiarato regge: se l'oggetto è l'intensità della contestazione, il volume di artefatti avversariali ben formati è un indicatore parziale ma operativamente definito. Non regge come misura di qualità argomentativa, che l'indice non cattura e non pretende di catturare.

L'unica eccezione è il componente delle assunzioni, modulato dalla gravità media delle assunzioni emerse. La gravità è l'unico segnale di qualità che entra nel numero, ed entra attraverso un canale con la sua vulnerabilità, come registra il §7.

Secondo: il proxy è *onesto* in un senso specifico e verificabile. È deterministico, scomponibile e versionato: due persone qualsiasi che lo calcolino dalla stessa sfida ottengono lo stesso numero, e possono attribuire quel numero alle sue parti.

L'onestà di un proxy non è la validità delle inferenze che se ne traggono (Messick, 1995). La prima la garantiamo nel §4, la seconda la differiamo nel §8. Chiamare il Δ-CSI «un proxy onesto» è un'affermazione sulla sua costruzione, non sul suo potere predittivo.

3.4 IL NOME: COSA MARCA IL DELTA

Il nome dell'indice si espande in *Cognitive Sovereignty Index, delta*. Un paper che introduce l'indice non può sottrarsi al dire cosa il delta marchi: la domanda è la prima che un referee solleverà, ed è sollevata contro l'oggetto stesso del paper. Fissiamo qui la risposta come convenzione di denominazione, enunciata anziché derivata.

Il delta marca il livello di analisi. Distingue una grandezza a *livello-sessione* — la pressione applicata all'interno di una singola analisi di una singola decisione — dall'indice a *livello-organizzazione*: il CSI di De Luca et al. (2026), che valuta la postura cognitiva stabile di un'istituzione.

Il delta non è una differenza temporale: non è un cambiamento dell'indice tra due istanti. E soprattutto non è una differenza nel giudizio del decisore, che è `decision_changed` e che l'indice esclude (§3.2). È un marcatore di livello. Il Δ-CSI sta al CSI organizzativo come una singola misurazione sta a una caratteristica stabile, e il delta registra che i due operano a livelli diversi senza alcun ponte formalizzato tra loro (§2).

Adottiamo la convenzione esplicitamente perché nessun documento precedente del corpus ha definito il delta, e un indice non può lasciare inspiegato il proprio nome. Dove un lettore si aspetta un operatore di differenza, gli chiediamo di leggere invece un indice di livello. È una convenzione di denominazione, e la marchiamo come tale anziché travestirla da risultato.

3.5 IL COSTRUTTO NEL QUADRO PIÙ AMPIO

Entro il quadro di *Sovranità cognitiva* (Canepa, 2026a), il Δ-CSI è descritto come un *aggancio parziale* del costrutto: una grandezza osservabile tra quelle attraverso cui la sovranità cognitiva diventa misurabile. Non è la sua misura complessiva, e non è un predittore della correttezza della decisione.

Questo paper è coerente con quella collocazione e non la estende. Il Δ-CSI àncora una dimensione — l'osservabilità della sfida applicata — a un livello, la sessione. Non misura la sovranità cognitiva: misura una cosa che un processo decisionale sovrano fa, una volta, e rende quella cosa verificabile.

SEZIONE 4

Definizione operativa

Questa sezione dà all'indice la sua forma esatta.

Le definizioni sono normative. Sono fissate, prima di ogni ulteriore prosa, in una specifica di accompagnamento: la `Δ-CSI Specification v1.0`, che vincola la versione di calcolo `csi-v1`. Il resoconto qui è l'enunciato leggibile di quel contratto, e dove i due si leggono insieme la specifica prevale. L'intento di progetto è che le formule qui sotto possano essere implementate e verificate senza ricorso ad alcuna altra fonte.

4.1 L'INPUT: UNA SFIDA STRUTTURATA

L'indice è calcolato su una **sfida**: il prodotto strutturato dell'analisi del contraddittore di una decisione.

Ai fini dello scoring, una sfida si riduce a quattro conteggi e a uno scalare. I conteggi sono il numero di assunzioni implicite emerse — ciascuna con una gravità intera opzionale su scala 1–5 — il numero di test di falsificazione, il numero di domande d'onere della prova e il numero di fonti citate. Lo scalare è un singolo valore di divergenza in [0, 1], derivato dal testo della tesi e dal testo del contro-scenario. Lo scorer consuma solo questi: i conteggi, le gravità delle assunzioni e lo scalare di divergenza. Non legge il testo libero della sfida al di là del conteggio.

Questo confine è ciò che rende riproducibile la parte quantitativa, mentre la parte narrativa — prodotta da un modello linguistico — non lo è (§5). È anche il senso preciso in cui l'indice misura la *struttura* che la pipeline ha prodotto anziché la *sostanza* degli argomenti (§3.3).

4.2 I CINQUE COMPONENTI

Il Δ-CSI è la somma di cinque contributi indipendenti. Nelle formule, `norm(n) = min(n, 5)/5` è il normalizzatore di conteggio con tetto, `severity_avg ∈ [1,5]` la gravità media delle assunzioni, `d ∈ [0,1]` la divergenza; i pesi core sono (25, 25, 20, 10, 20).

COMPONENTE	FORMULA	MASSIMO
assunzioni	`25 · norm(n) · (severity_avg / 5)`	25
test di falsificazione	`25 · norm(n)`	25
domande d'onere della prova	`20 · norm(n)`	20
fonti citate	`10 · norm(n)`	10

COMPONENTE	FORMULA	MASSIMO
divergenza	`20 · d`	20

I cinque si sommano in un punteggio in [0, 100]. I componenti sono additivi, senza termini incrociati: ogni punteggio è quindi scomponibile, e dato un Δ-CSI si può sempre dire quanti punti vengono da ciascuna dimensione. È una proprietà deliberata. Un aggregato opaco vanificherebbe la verificabilità, che è l'intero scopo del misurare la sfida.

Due scelte di modellazione, qui, sono stipulazioni e non risultati, e il paper le marca come tali.

Il Δ-CSI è un composito *formativo*: i suoi componenti costituiscono l'indice per definizione, invece di riflettere un'unica variabile latente che li causi — come farebbero invece gli item di una scala psicometrica riflessiva. Da ciò discendono due obblighi, e il paper non ne assolve nessuno.

Il primo riguarda la scelta dei componenti. Che siano esattamente questi cinque andrebbe argomentato come copertura dell'intero dominio dell'intensità della sfida. È una pretesa di copertura che il paper asserisce ma non dimostra: i cinque, in effetti, rispecchiano lo schema di output fisso del contraddittore.

Il secondo riguarda la forma. Quella additiva e compensativa è a sua volta un'assunzione. Una somma lineare permette a un risultato forte su quattro componenti di riscattarne uno debole sul quinto; un costrutto congiuntivo lascerebbe invece che un canale debole ponesse un tetto all'intero indice. Sia la scelta dei componenti sia la forma additiva figurano tra le debolezze dichiarate (§7.2).

Ciò qualifica anche l'inquadramento in termini di validità di costrutto del §2. Un composito formativo si valida soprattutto attraverso la copertura di contenuto e le relazioni di criterio, non attraverso la logica di coerenza interna della tradizione riflessiva di cui Cronbach, Meehl e Messick si occupano per lo più. È la ragione per cui l'evidenza rilevante sono i test convergenti e a gruppi noti della roadmap (§8.2), non le statistiche di coerenza interna.

4.3 DUE REGOLE CHE TENGONO ONESTO IL NUMERO

Due regole di costruzione svolgono il lavoro portante, e ciascuna risponde a un modo specifico in cui il numero potrebbe essere corrotto.

Il tetto, contro l'inflazione per volume. Ogni componente basato sul conteggio satura a cinque: `norm(5) = norm(50) = 1,0`. Il valore marginale di una sesta assunzione, o di un decimo test di falsificazione, è esattamente zero. Dieci obiezioni superficiali non superano cinque profonde.

La regola toglie l'incentivo a fabbricare volume. Obbliga a cercare un punteggio più alto nella sostanza della sfida — un contro-scenario più distante, o assunzioni giudicate più gravi — invece che nella lunghezza degli elenchi. Anche questa seconda via passa però per il canale della gravità auto-dichiarata, la cui vulnerabilità è registrata al §7.2. Un lettore

non dovrebbe perciò mai inferire un conteggio esatto da un componente saturo: un contributo massimo di `falsifications` certifica «almeno cinque», non un totale.

Il vincolo di somma a cento, contro l'inflazione occulta del tetto. I cinque pesi sommano a cento: ogni peso è la quota massima che il suo componente può apportare, e il punteggio massimo possibile è cento.

La somma a cento è essa stessa una convenzione stipulata, imposta programmaticamente. Ciò che ne segue, dati pesi fissi, è che nessuna ridistribuzione dei pesi può gonfiare occultamente il tetto.

Una configurazione di dominio (un «Pack») può ridistribuire i pesi — un contesto clinico o di due-diligence legale potrebbe alzare il peso delle fonti a spese della divergenza — ma solo alla condizione che la somma resti cento. Il vincolo proibisce di alzare il tetto di un componente senza abbassare quello di un altro. Le fasce di interpretazione (sotto) non sono mai spostabili. La conseguenza, sviluppata al §6, è che punteggi prodotti con pesi diversi non sono direttamente confrontabili e vanno tenuti in coorti di calibrazione separate.

4.4 GRAVITÀ DELLE ASSUNZIONI: L'UNICO CANALE DI QUALITÀ

Il componente delle assunzioni è l'unico modulato da un segnale di qualità anziché da un conteggio puro. Il suo moltiplicatore `severity_avg / 5` varia da 0,2 (tutte le assunzioni a gravità 1) a 1,0 (tutte a gravità 5): due sfide con lo stesso numero di assunzioni ma gravità opposta producono punteggi diversi.

Due convenzioni di default governano i dati mancanti. Una singola assunzione senza gravità dichiarata conta come 3, il punto medio neutro. Ma una sfida che non dichiara *alcuna* gravità riceve `severity_avg = 5` — il massimo — per retrocompatibilità con la versione di calcolo pre-gravità.

La seconda convenzione è un difetto noto, che segnaliamo qui e riprendiamo al §7: premia l'omissione rispetto alla dichiarazione onesta di valori medi. È un incentivo perverso, che le pipeline di produzione sono indirizzate a evitare dichiarando sempre le gravità, e che il paper si rifiuta di nascondere.

4.5 DIVERGENZA: LESSICALE E SEMANTICA

Il componente di divergenza misura quanto il contro-scenario si allontana dalla tesi, su scala [0, 1]. Lo fa con uno di due metodi, che misurano *proprietà diverse* e non sono perciò mutuamente confrontabili.

Il metodo lessicale è il complemento di Jaccard sugli insiemi di token normalizzati: `1 − |T(τ) ∩ T(κ)| / |T(τ) ∪ T(κ)|` (Jaccard, 1912). È deterministico, offline e senza dipendenze, e serve da fallback garantito. Registra però come divergenti una tesi e un contro-scenario che concordano nella sostanza ma differiscono nelle parole — una distorsione rispetto alla proprietà d'interesse.

Il metodo semantico mappa la similarità coseno tra i vettori di embedding dei due testi in una distanza: `(1 - cos)/2`, nella tradizione dello spazio vettoriale di Salton, Wong e

Yang (1975). È più fedele al costrutto, al costo di una dipendenza da un encoder esterno.

Quale metodo sia stato usato lo decide la configurazione ed è, cosa cruciale, *registrato su ogni punteggio*. Se un encoder semantico è richiesto ma fallisce a runtime, il calcolo ripiega sul metodo lessicale ed etichetta il record di conseguenza: punteggi di tipo genuinamente diverso non sono mai fusi in silenzio.

Le due misure di divergenza non sono intercambiabili nemmeno a parità di valore numerico. Il punto ricorre come vincolo di confrontabilità nel §6.

4.6 FASCE DI INTERPRETAZIONE E VERSIONING

Il punteggio si legge attraverso tre fasce fisse — *marginale* [0, 33), *moderata* [33, 66), *sostanziale* [66, 100] — identiche in ogni dominio e configurazione.

La fissità è deliberata. Se le installazioni fossero libere di spostare le soglie, la parola «moderata» perderebbe il suo significato condiviso, e lo stesso numero racconterebbe storie diverse a interlocutori diversi. L'interpretazione canonica è sempre un'affermazione sulla *pressione sulle assunzioni della decisione* — «il contraddittore ha applicato pressione moderata» — mai sul merito della decisione o su un cambiamento nella mente del suo autore.

Il calcolo è versionato. La versione corrente (`csi-v1`) differisce dalla precedente (`csi-v0`) in un solo aspetto: la modulazione per gravità del componente delle assunzioni.

Ogni record porta la propria versione. E poiché una sfida con gravità dichiarata riceve un punteggio diverso sotto le due versioni, i punteggi di versioni diverse sono tenuti in coorti separate e mai mescolati (§6). Qui il versioning non è zavorra burocratica: è ciò che permette alla definizione dell'indice di cambiare senza invalidare in silenzio i punteggi calcolati sotto la definizione precedente.

4.7 UN ESEMPIO LAVORATO

La costruzione è concreta su un singolo caso, riprodotto dall'implementazione di riferimento e verificabile a mano.

Una sfida fa emergere tre assunzioni senza gravità dichiarate (quindi `severity_avg = 5,0`), due test di falsificazione, due domande d'onere, una fonte, e una divergenza lessicale di 0,9615. I componenti sono `25·(3/5)·(5/5) = 15,00`, `25·(2/5) = 10,00`, `20·(2/5) = 8,00`, `10·(1/5) = 2,00`, e `20·0,9615… = 19,23`, sommando a **54,23** — nella fascia *moderata*. Il numero è prodotto deterministicamente dallo scorer di riferimento: dati gli stessi cinque input, restituisce 54,23 ogni volta.

Questo esempio canonico è mantenuto per continuità con la documentazione di riferimento, ma gira deliberatamente su due delle strade più deboli dell'indice. Lo segnaliamo anziché nasconderlo. Le assunzioni non portano *alcuna gravità dichiarata*, quindi il componente delle assunzioni gira al massimo di retrocompatibilità (`severity_avg = 5,0` — l'incentivo all'omissione del §4.4 e §7.2). E la divergenza è il fallback *lessicale* di Jaccard, il meno valido dei due metodi (§4.5).

Un'illustrazione metodologicamente più pulita dichiarerebbe gravità medie e userebbe la divergenza semantica. Il numero differirebbe, ma il punto dell'esempio lavorato — l'aritmetica e la sua riproducibilità a mano — sarebbe invariato.

La narrativa che una sfida reale avvolgerebbe attorno a questi input — il testo effettivo delle assunzioni, lo scenario, le domande — è generata da un modello linguistico e *non* è deterministica. Solo la parte quantitativa riproduce esattamente. Quella distinzione, tra un numero riproducibile e un testo non riproducibile, è l'oggetto della prossima sezione.

SEZIONE 5

La pipeline di misura e la separazione provenienza/scoring

Il Δ-CSI non nasce dal nulla. È l'ultimo anello di una catena, e sia le sue garanzie sia i suoi limiti sono leggibili solo rispetto al processo che ne produce gli input.

Il contraddittore opera come tre passate sequenziali e isolate, orchestrate deterministicamente. La prima (Tesi) estrae la lettura dominante della decisione — ciò che un buon analista concluderebbe — e non è essa stessa avversariale.

La seconda (Red Team) è la sfida vera e propria. Fa emergere le assunzioni implicite, costruisce il contro-scenario, propone i test di falsificazione e pone le domande d'onere della prova. Trae il proprio mandato, la propria tassonomia di bias e i propri template di falsificazione dalla configurazione di dominio risolta.

La terza (Sintesi) calibra il grado di confidenza alla luce di tesi e sfida, dichiara i gap residui e associa le fonti alla loro provenienza. Non ha mandato di concludere o raccomandare.

Due caratteristiche della pipeline incidono direttamente sul significato dell'indice.

La prima è l'*isolamento tra modelli nelle passate*. Le passate possono girare su provider di modello diversi: la lettura dominante di una decisione non viene così attaccata dallo stesso modello che l'ha prodotta. I provider effettivamente usati sono registrati con l'output.

La motivazione è una modalità di guasto documentata nella letteratura sul debate multi-agente (Wynn, Satija, & Hadfield, 2025; Wu, Li, & Li, 2025): cornici avversariali imposte a un singolo modello tendono a produrre ragionamento correlato e disaccordo inscenato anziché genuino. Isolare le passate tra provider è un tentativo strutturale di de-correlare i punti ciechi sistematici, cosicché la sfida che il Δ-CSI misura sia più probabilmente reale.

È una mitigazione, non una prova. Se riesca è una delle cose che una validazione eventuale dovrebbe stabilire.

Comporta anche un costo, che il paper non nasconde. Far girare le passate su provider diversi inietta varianza tra modelli proprio nei conteggi e nelle gravità da cui l'indice è calcolato. La stessa decisione può così produrre componenti diversi — e quindi un Δ-CSI diverso — sotto instradamenti diversi dei provider. La conseguenza sull'affidabilità è ripresa sotto ed elencata come debolezza (§7.2).

La seconda caratteristica, e la più importante per l'interpretabilità dell'indice, è la **separazione della provenienza dallo scoring**. Il companion di costrutto la tratta come uno dei quattro invarianti del contraddittore (Canepa, 2026a, §4.4); qui la riformuliamo nella sua forma di misura.

Un contraddittore dotato di retrieval potrebbe introdurre una circolarità. Le evidenze trovate durante l'analisi potrebbero simultaneamente plasmare il contenuto della sfida *e* gonfiarne l'intensità misurata. Un Δ-CSI alto diventerebbe allora un artefatto di quante fonti

si sono rese disponibili, anziché di quanta pressione il ragionamento abbia effettivamente ricevuto, e l'indice cesserebbe di misurare la proprietà che nomina.

La separazione preclude questo. Le fonti sono recuperate e registrate come provenienza, ma i componenti del Δ-CSI sono calcolati dalla *struttura logica* della sfida — assunzioni, test di falsificazione, domande d'onere, divergenza dello scenario — non dal volume o dalla qualità del materiale recuperato. Il conteggio delle fonti entra come un componente separato, con il proprio peso limitato (dieci su cento), e non retroagisce sugli altri. È la condizione sotto cui il Δ-CSI resta interpretabile come intensità della sfida, anziché come misura mascherata della ricchezza del retrieval. I quattro componenti non-fonte sono isolati dal contesto informativo, e il componente delle fonti è limitato a dieci su cento: la ricchezza del retrieval può spostare il punteggio, ma non dominarlo.

Un indice confuso con l'abbondanza di fonti non potrebbe essere calibrato affatto nel senso del forecasting (Tetlock & Gardner, 2015), perché misurerebbe una proprietà del corpus anziché del ragionamento.

La pipeline produce quindi due output qualitativamente diversi, e la distinzione è essenziale a ciò che il paper afferma.

La narrativa — tesi, scenario, assunzioni, domande — è generata da un modello linguistico e non è riproducibile: la stessa decisione analizzata due volte produce testi diversi. Il punteggio, invece, è calcolato da una funzione deterministica sulla struttura estratta, ed *è* riproducibile: gli stessi componenti danno sempre lo stesso numero.

Quando il paper parla di riproducibilità (§4.7, §7), la riferisce allo scorer, non alla pipeline. Un test di regressione dell'indice deve fissare i componenti come input, e non attendersi determinismo dallo stadio narrativo. Confondere i due livelli sarebbe un'affermazione che il sistema non può mantenere, e non la facciamo.

La conseguenza onesta è che la *misura*, dall'inizio alla fine, non è riproducibile nel modo in cui lo è lo *scorer*. Poiché conteggi, gravità e divergenza sono estratti da un modello stocastico, la stessa decisione analizzata due volte riceverà in generale un Δ-CSI diverso. Il determinismo che il paper garantisce è quello dell'aggregazione a partire da componenti fissi: una proprietà dell'aritmetica, non della misura.

Quanto il punteggio end-to-end vari tra esecuzioni e tra provider — la sua affidabilità test-retest e inter-model — non è quantificato qui. È in testa alla roadmap di validazione del §8.2, ed è tra le debolezze dichiarate del §7.2.

SEZIONE 6

Persistenza e disciplina di calibrazione

Un indice calcolato una volta e scartato misura un momento. Un indice persistito e rivisitato può, in linea di principio, essere validato.

Il Δ-CSI è scritto, per ogni sessione, in un **Sovereign Decision Ledger** per-tenant e local-first. Il ledger registra la decisione, la sfida completa, l'indice con la sua scomposizione per componente e la versione, e — in differita, quando diventa noto — l'esito. L'esito ha due campi: un `outcome_label` (corretta, errata, mista o ignota) e il booleano `decision_changed`.

Il ledger è per-tenant e isolato, e il suo record accumulato di sfide ed esiti è il substrato su cui opera la calibrazione. Ai fini di questo paper conta non come asset commerciale, ma come il meccanismo che rende la validità empirica *costruibile nel tempo* anziché assumibile in anticipo.

È necessario essere esatti su cosa significhi «calibrato» per il Δ-CSI: la parola fa un lavoro più stretto di quanto suggerirebbe la letteratura del forecasting da cui è presa. La calibrazione qui è una *disciplina*, non uno stato che l'indice ha raggiunto. Ha tre regole.

Coorti. Le statistiche di calibrazione sono calcolate per coorte, dove una coorte è la tripla (versione del calcolo, metodo di divergenza, pesi).

Score prodotti sotto versioni diverse, metodi di divergenza diversi o pesi diversi non sono confrontabili, e non sono mai messi insieme. Un Δ-CSI di 54 sotto divergenza lessicale e un Δ-CSI di 54 sotto divergenza semantica non affermano la stessa cosa sull'intensità della sfida: la loro identità numerica è accidentale (§4.5). Mescolarli produrrebbe statistiche tratte da misurazioni incommensurabili, che non possono essere interpretate come campioni di una sola variabile.

Un pavimento puramente descrittivo. Al di sotto di un minimo di cinque record con esito in una coorte, l'output di calibrazione è esplicitamente marcato come puramente descrittivo, e nessuna conclusione sulla validità dell'indice è consentita.

Non è una soglia di significatività: è un pavimento operativo di minimo buon senso. Superarlo non licenzia alcuna pretesa di validità, e anche al di sopra l'evidenza resta descrittiva e locale al tenant. Il gate esiste per prevenire un errore specifico: presentare il Δ-CSI come validato empiricamente prima che alcun esito si sia accumulato.

Una statistica locale e descrittiva. Al di sopra del pavimento, la calibrazione può riportare un `hit_rate`: la proporzione di decisioni *cambiate* con esito definitivo che si sono rivelate corrette.

La definizione è deliberatamente stretta. Condiziona su `decision_changed`, perché la domanda che la calibrazione può onestamente porre al ledger è se il contraddittore, *quando ha effettivamente spostato una decisione*, l'abbia spostata bene.

Il `hit_rate` può essere indefinito anche al di sopra del pavimento: se nessuna decisione nella coorte è mai stata cambiata, il denominatore è zero. Non è un'anomalia, ma un cor-

retto resoconto dell'assenza dell'evidenza rilevante.

Dove definito, il `hit_rate` è per-tenant, descrittivo e locale. Non è una stima causale, e un valore alto in un dominio non implica nulla su un altro.

Due confini dell'apparato di calibrazione meritano di essere enunciati, perché la loro omissione sopravvaluterebbe ciò che il ledger consegna.

Primo, l'apparato come specificato non mette in relazione il *valore* del punteggio con gli esiti. Nessuna delle sue statistiche, `hit_rate` incluso, è funzione della grandezza del Δ-CSI: l'apparato accumula l'evidenza che una futura analisi di validità richiederebbe, senza esso stesso costituire un test della validità dell'indice.

Secondo, il pavimento a livello di engine di cinque esiti è assai più debole della potenza statistica che uno studio confermativo esigerebbe. Il protocollo companion (Canepa, 2026b) fissa quell'asticella in modo esplicito: un pavimento di almeno venti casi con punteggio per orizzonte, per qualsiasi test di calibrazione confermativo.

È importante che le due soglie non siano confuse. Superare il gate di cinque esiti del prodotto autorizza un tenant a *leggere* statistiche descrittive, non a *pretendere* validazione. La distanza tra questi due atti è la distanza tra questo paper e il suo successore empirico.

Al momento della scrittura, quella distanza non è stata colmata. Nessuna coorte per-tenant ha raggiunto nemmeno il pavimento descrittivo di cinque esiti con decisori reali: il volano è specificato e strumentato, ma non ancora messo in moto dall'uso reale.

Il paper riporta questo come un fatto sullo stato dell'evidenza, nello stesso registro con cui il protocollo companion riporta lo stato pre-esito del suo pilot sigillato. La dichiarazione onesta è parte del contributo: un paper di misura che oscurasse il vuoto del proprio ledger di esiti commetterebbe, in miniatura, la sopravvalutazione a cui l'intero programma esiste per resistere.

SEZIONE 7

Auto-contraddittorio: proprietà, difese e debolezze dichiarate

Un indice il cui scopo è misurare l'intensità dell'auto-contestazione deve essere disposto a contestare sé stesso, e la disposizione va dimostrata anziché asserita.

Questa sezione rivolge il metodo sul proprio oggetto. Enuncia prima le difese strutturali che fanno fallire l'indice in modo visibile anziché adulare in silenzio; poi, senza mitigazione, le debolezze che sopravvivono a quelle difese.

7.1 TRE DIFESE FAIL-CLOSED

Il sistema implementa tre protezioni strutturali contro l'output compiacente, strutturali nel senso che sono controlli a livello di codice che scattano indipendentemente da ciò che il modello produce.

La prima è il *Contradiction Floor*. Una sfida è ammissibile solo se contiene almeno un'assunzione, almeno un test di falsificazione, almeno una domanda d'onere e un controscenario non vuoto.

Se il floor non regge, il sistema esegue esattamente un retry affilato. Se non regge ancora, la pipeline solleva un errore controllato e non emette nulla. Quando non c'è alcuna sfida genuina da produrre, il sistema fallisce in modo visibile invece di ripiegare su un resoconto rassicurante: una sfida vuota diventa un errore, non un risultato.

Il floor è un pavimento e non un soffitto, e la sua soglia è deliberatamente bassa — il che è a sua volta una delle debolezze dichiarate sotto. Ma la sua funzione è categorica, ed è la prima linea contro la modalità di guasto che l'intero indice esiste per rilevare.

La seconda è la *provenienza a enumerazione chiusa*. Ogni fonte deve dichiarare una provenienza da un insieme chiuso di cinque valori. Senza retrieval disponibile, la sintesi è vincolata a restituire un elenco di fonti vuoto anziché inventare citazioni.

Una fonte non può materializzarsi dal nulla: o ha un'origine dichiarata, o non entra nel conteggio. Questo chiude il fallimento più insidioso di un sistema che cita, cioè la fabbricazione di riferimenti plausibili ma inesistenti.

La terza è la *calibrazione onesta sotto soglia* (§6): al di sotto di cinque esiti il sistema dichiara insufficienti le proprie statistiche, anziché simulare una validazione che non ha. La difesa è di nuovo quella del fallimento visibile anziché della degradazione silenziosa. L'output corretto di un ledger vuoto è una dichiarazione di insufficienza, non un numero rassicurante.

7.2 DEBOLEZZE DICHIARATE

Le seguenti debolezze sono aperte. Sono enunciate qui perché una misura la cui validità poggia in parte sull'onestà del proprio auto-resoconto non può trattare i propri limiti come una minaccia da gestire. In una lettura di due-diligence, il valore dell'elenco non è l'assenza di voci: è il fatto che il sistema le dichiari esso stesso.

Il nome e la scala invitano a un errore di categoria. L'etichetta «Cognitive Sovereignty Index» e la scala 0–100 invitano a leggere il numero come una misura di correttezza. Non lo è (§3.2).

Il rischio è comunicativo, e non è neutralizzato dall'implementazione attuale. La mitigazione è la disciplina definitoria — l'indice deve viaggiare sempre con un'affermazione esplicita di ciò che misura — non una garanzia strutturale. È il primo rischio elencato dal paper stesso, ed è la ragione per cui il §3 precede il §4.

Tre componenti misurano solo la quantità. Test di falsificazione, domande d'onere e fonti sono conteggi puri (§4): cinque elementi banali valgono quanto cinque incisivi, e la qualità degli argomenti non entra nel numero. Solo il componente delle assunzioni porta un segnale di qualità.

Il floor è permissivo. Il Contradiction Floor richiede un elemento per campo: una sfida con una sola assunzione ovvia, un test banale, una domanda e uno scenario di una riga lo supera. La difesa contro le sfide vuote è reale, ma la sua soglia è bassa. Una sfida sottile-ma-non-vuota passa e si presenta con un punteggio marginale — da leggere con sospetto (§3.2), non come rassicurazione.

Una sintesi senza fonti passa senza allarme. Un elenco di fonti vuoto supera il floor e la validazione dello schema. La sintesi è salvata e riceve un punteggio, con il componente delle fonti che contribuisce con zero, e nulla segnala esplicitamente l'assenza al lettore. Il comportamento è difendibile — un'assenza onesta è meglio di una citazione fabbricata — ma lo zero silenzioso è un gap dichiarato.

Sotto cinque esiti, nulla è validato empiricamente. Finché una coorte non accumula cinque esiti definitivi, la calibrazione è descrittiva e nessuna conclusione di validità è licenziata (§6). È un limite di evidenza, riformulato come debolezza perché è di routine dimenticato nella presentazione di nuove metriche.

L'unico segnale di qualità è auto-dichiarato dal modello. La gravità delle assunzioni — l'unico canale di qualità che entra nel numero — è dichiarata dallo stesso modello che genera la sfida.

Il tetto anti-inflazione governa i conteggi, non le gravità. Un modello che sovrastima sistematicamente la gravità gonfia il componente dal peso massimo, e nessuna difesa strutturale lo intercetta.

È la debolezza aperta più consequenziale dell'indice, e ha la struttura di una patologia di misura classica. Una volta che una grandezza calcolata dalle dichiarazioni di gravità di un modello è usata per giudicare l'output di quel modello, la dichiarazione è spinta a servire il punteggio anziché a descrivere l'assunzione. Nella formulazione originale di Goodhart (1975), è il collasso di una regolarità statistica sotto controllo. Nella sua generalizzazione or-

mai standard, è la tendenza di una misura a cessare di essere una buona misura una volta che diventa un obiettivo (Strathern, 1997, sulla scia di Hoskin).

L'indice è, sotto questo aspetto, aggirabile per costruzione attraverso il canale della gravità, e lo dice. La vulnerabilità è poi un caso specifico di una generale e documentata. Usare un modello linguistico per valutare l'output di un modello linguistico importa i bias noti dell'impostazione LLM-come-giudice nell'unico canale che porta informazione di qualità. Il principale è la self-preference: un modello valuta favorevolmente il proprio output (Zheng et al., 2023).

Omettere le gravità dà un punteggio più alto che dichiararle onestamente. Per retrocompatibilità, una sfida che non dichiara alcuna gravità riceve la gravità media massima. Una sfida che le ometta tutte ottiene così un punteggio più alto di un'identica con gravità medie dichiarate onestamente (§4.4).

È un incentivo perverso noto, accettato come costo della compatibilità di versione. Le pipeline di produzione sono istruite a dichiarare sempre le gravità; il difetto è dichiarato anziché riparato.

La misura end-to-end ha affidabilità non quantificata. Il determinismo che il paper garantisce è quello dello scorer, dati componenti fissi. I componenti stessi sono estratti da un modello stocastico: la stessa decisione ri-analizzata riceverà in generale un Δ-CSI diverso, e l'instradamento a isolamento di modello del §5 aggiunge varianza tra provider al di sopra.

L'affidabilità test-retest e inter-model del punteggio end-to-end — e dunque la stabilità della fascia in cui una decisione cade — non è misurata in questo paper. Poiché l'affidabilità precede la validità, questo è il più consequenziale dei gap pre-empirici, e apre la roadmap del §8.2.

I pesi e la forma additiva sono stipulati, non derivati. I pesi core (25/25/20/10/20) sono un giudizio di default sul contributo relativo delle cinque dimensioni, non una grandezza stimata da dati o elicitata da esperti. E l'aggregazione additiva e compensativa è un'assunzione di modellazione, non una proprietà dimostrata del costrutto (§4.2). Un'analisi di sensibilità sui pesi — con quale frequenza le assegnazioni di fascia cambino sotto ridistribuzioni ragionevoli — limiterebbe l'arbitrarietà, e non è ancora eseguita.

Il componente delle assunzioni non è monotono nella sfida genuina. Poiché il componente moltiplica un conteggio con tetto per la gravità *media*, aggiungere un'ulteriore assunzione genuina ma più mite può *abbassare* il punteggio. Cinque assunzioni a gravità 5 danno il massimo di 25,00, ma una sesta assunzione genuina a gravità 1 fa scendere il componente a 21,67: il conteggio è già saturo, e la media cala.

Una sfida in senso stretto più ampia e onesta può così ottenere un punteggio più basso. La non-monotonicità crea anche un incentivo a *sopprimere* le assunzioni miti per tenere alta la media — un incentivo indipendente dal tetto anti-verbosità.

È una proprietà reale della formula `csi-v1` attuale, dichiarata qui anziché portata in silenzio. Un'aggregazione monotona, per esempio una somma pesata per gravità sotto lo stesso tetto, è una correzione candidata per una futura versione del calcolo.

Sotto un encoder reale, la divergenza semantica non può raggiungere il suo tetto. La mappatura coseno `(1 - cos)/2` assume che il coseno copra [−1, 1]. Ma i coseni degli embedding di frase per coppie in linguaggio naturale sono tipicamente positivi e alti: la divergenza semantica occupa all'incirca la parte bassa di [0, 1], e il suo componente contribuisce realisticamente solo pochi dei suoi venti punti. Le fasce fisse finiscono così per suddividere intervalli effettivi strutturalmente diversi tra la coorte lessicale e quella semantica, e la stessa sfida, valutata con i due metodi, può cadere in fasce diverse.

Le coorti tengono i due fuori da un unico pool di calibrazione (§6), ma non fanno arrivare la scala semantica in alto quanto quella lessicale. Una mappatura del coseno calibrata sulla distribuzione sarebbe la correzione di principio.

Diverse di queste voci — la divergenza semantica come default disponibile, la modulazione per gravità del componente delle assunzioni — erano esse stesse la risoluzione di voci precedenti in questo stesso auto-contraddittorio. Altre — il gap di affidabilità, la non-monotonicità, i pesi stipulati — sono emerse rivolgendo una review avversariale su questo paper stesso.

È la dinamica intesa: il registro delle debolezze è un backlog vivo, non una confessione appesa una volta. Ciò che l'elenco dimostra non è che l'indice sia privo di difetti, perché non lo è. Dimostra che i suoi difetti sono fatti emergere dalla stessa disciplina avversariale che l'indice applica alle decisioni che valuta. Un fornitore che li presentasse come già risolti affermerebbe più di quanto il sistema affermi di sé.

SEZIONE 8

Limiti e una roadmap di validazione pre-registrata

Le affermazioni del paper sono delimitate, e il confine è il punto.

Ciò che è stabilito qui è un costrutto e un metodo di misura: il Δ-CSI è definito, la sua aritmetica è deterministica e riproducibile, le sue difese e debolezze sono specificate, e le sue regole di confrontabilità sono fissate.

Ciò che *non* è stabilito è alcuna validità empirica, cioè che un dato livello di Δ-CSI corrisponda a qualcosa nelle decisioni che valuta. Quell'affermazione è differita, e questa sezione ne enuncia i termini con precisione sufficiente perché se ne debba rendere conto.

8.1 CIÒ CHE L'INDICE NON È: CORROBORAZIONE

Un lettore formato nella filosofia della scienza udirà, in una frase come *l'intensità con cui una decisione è stata contestata*, un'eco della corroborazione popperiana e della tradizione del severe testing che la formalizza (Popper, 1959; Mayo, 1996).

Vale la pena affrontare l'eco direttamente, perché la somiglianza è in parte reale e in parte una trappola. È reale nello spirito: entrambe le tradizioni tengono al confrontare un'affermazione con ciò che potrebbe rovesciarla, anziché con ciò che la conforterebbe.

È una trappola se letta come identità, e la ragione è precisa. Il severe testing è definito dal *potere probante* di un test: un test è severo nella misura in cui avrebbe probabilmente esposto un errore, se ce ne fosse stato uno. Il Δ-CSI non misura nulla del genere. Come enunciano chiaramente il §3.3 e il §7.2, conta i test di falsificazione senza graduare se alcuno sia probante: cattura la *quantità e l'intensità della contestazione applicata*, non la *severità* di alcun test nel senso di Mayo. La parola «gravità/severità» all'interno dell'indice si riferisce peraltro a un'altra cosa ancora — la gravità auto-valutata di un'*assunzione* (§4.4) — e non va confusa con la severità del test.

La distinzione decisiva, tuttavia, è temporale, ed è ciò che rende «non corroborazione» vero anziché soltanto asserito. La corroborazione giudica un'ipotesi per i test che ha *superato*, dopo che gli esiti sono noti. Il Δ-CSI assegna un punteggio a una decisione per le sfide *applicate ad essa*, prima che alcun esito sia noto e a prescindere dal fatto che la decisione sia poi confermata.

L'indice non è perciò una misura di corroborazione sotto nuovo nome. È un audit della contestazione applicata ex ante, e la sua stessa validazione (sotto) passa per gli esiti decisionali in un ledger, non per la sopravvivenza di ipotesi sotto esperimento. Prende in prestito lo spirito avversariale della tradizione del severe testing; non ne prende in prestito, e non ne pretende, la misura di severità probante.

8.2 LA ROADMAP DI ACCUMULO DELL'EVIDENZA

La validazione è trattata come una questione di progettazione del processo nel tempo, pre-registrata anziché asserita, e procede attraverso soglie che non vanno confuse.

Prima di qualsiasi delle soglie dipendenti dall'esito qui sotto, c'è un lavoro di misura che non richiede alcun esito, e questo paper lo differisce anziché eseguirlo. Due studi sono pre-registrati qui come passo immediato successivo.

Il primo è uno *studio di affidabilità*: un insieme fisso di decisioni fatto girare ripetutamente, su almeno due instradamenti di provider. Riporterebbe la varianza tra esecuzioni e tra modelli di ciascun componente e del punteggio totale, un coefficiente di affidabilità test-retest, l'accordo sull'assegnazione di fascia, e un errore standard di misura da apporre a ogni Δ-CSI riportato.

Il secondo è uno *studio di discriminazione*: un corpus etichettato di sfide genuine contro sfide deliberatamente simulate («sham»). È un test a gruppi noti che verifichi se l'indice le separi — proprio la capacità che il §2 indica come scopo dell'indice, e non ancora dimostrata.

Un controllo convergente appartiene allo stesso stadio, e non richiede esiti nemmeno esso: l'accordo tra il Δ-CSI e valutazioni indipendenti di esperti sull'intensità della sfida, sulle stesse sfide. Darebbe alla rete nomologica altrimenti solo-discriminante (§2) il suo primo nodo confermativo.

Nessuno di questi richiede che il ledger si riempia, né utenti vivi, né esiti decisionali. Sono la precondizione, non la conseguenza, del lavoro sugli esiti che segue, e la loro assenza è precisamente la ragione per cui questo paper è un contributo di definizione-e-metodo anziché una misura validata.

La prima soglia è il pavimento descrittivo dell'engine: cinque record con esito in una coorte, al quale un tenant può cominciare a *leggere* il record accumulato di punteggi ed esiti in modo descrittivo (§6).

Una lettura pragmatica — sufficiente a rilevare un segnale che valga la pena inseguire — richiede dell'ordine di venti-trenta record con un campo `decision_changed` definito. È una scala raggiungibile da un singolo utente serio in mesi, non un programma di ricerca.

L'evidenza investor-grade, negli obiettivi interni del corpus stesso, è un ordine di grandezza ancora diverso: centinaia di esiti definitivi per verticale. È nominata qui solo per marcare il divario tra il segnale descrittivo di un pilot e un'affermazione empirica difendibile, cosicché il primo non sia mai presentato come il secondo.

La seconda soglia appartiene al protocollo companion e valida una grandezza adiacente, che il paper è attento a non sovra-attribuire all'indice.

Calibrated Dissent (Canepa, 2026b) mette in atto un audit pre-registrato e a prova di senno di poi. Il suo oggetto è un pilot fisso e pubblicamente sigillato di dieci decisioni M&A europee ad alto rischio, sigillate a una singola data (20 giugno 2026), ciascuna con previsioni risolte a due orizzonti: sei-dodici mesi e diciotto-trentasei mesi.

Quel protocollo assegna un punteggio alle *previsioni* sigillate rispetto agli esiti, con regole di scoring proprie. Fissa un pavimento di potenza confermativa di almeno venti casi con

punteggio per orizzonte — soglia che il pilot fisso di dieci, per costruzione, non raggiunge. Ne seguono due punti di disciplina, entrambi portanti per l'onestà di questo paper.

Primo, in quel protocollo il Δ-CSI entra unicamente come *provenienza non conteggiata nel punteggio* e, al più, come covariata esplorativa. È registrato, non valutato, e nessuna ipotesi confermativa è enunciata su di esso.

Secondo, e di conseguenza, lo stadio a potenza piena del companion valida le *previsioni* del contraddittore, non il Δ-CSI in quanto tale. Un forecaster ben calibrato è evidenza sulla pipeline, non una validazione diretta dell'indice di intensità della sfida.

La via di validazione dell'indice stesso resta la calibrazione del ledger per-tenant del §6, cioè la relazione accumulata tra intensità della sfida ed esiti decisionali registrati. Quella via, al momento della scrittura, non ha ancora prodotto alcun esito.

8.3 CONFLITTO D'INTERESSE E LA MODALITÀ DI GUASTO DEL GAMING

Due dichiarazioni chiudono la sezione, nel registro che il protocollo companion adotta per gli stessi scopi.

La prima è un conflitto d'interesse, dichiarato anziché occultato. L'autore di questo paper è l'autore del paper di costrutto, il costruttore dell'implementazione di riferimento, e una parte con un interesse a che il metodo di misura sia credibile. Un lettore dovrebbe pesare le affermazioni del paper alla luce di ciò.

Le mitigazioni sono quelle disponibili a un singolo autore: una specifica aperta e irrevocabile (§9), la pre-registrazione dei test empirici nel companion, e l'intenzione di porre lo scoring degli esiti sotto risoluzione indipendente. Nessuna di queste rimuove il conflitto. Tutte lo rendono visibile e verificabile, che è il massimo che una dichiarazione onesta possa fare.

La seconda è la modalità di guasto principale dell'indice, già enunciata come debolezza (§7.2) e ripetuta qui come bersaglio di falsificazione pre-registrato: il *gaming*.

Si supponga che un deployment si scopra produrre punteggi Δ-CSI gonfiati senza alcun corrispondente aumento nella sostanza della sfida, con ogni probabilità attraverso il canale della gravità auto-dichiarata. Sarebbe evidenza che l'indice è diventato un obiettivo e ha cessato di essere una buona misura (Strathern, 1997). La condizione è enunciata in anticipo, come la disciplina della pre-registrazione richiede, cosicché la sua comparsa successiva conti come disconfermante anziché essere spiegata via.

SEZIONE 9

La specifica come standard aperto

La definizione esposta in questo paper è offerta a termini aperti, cosicché altri possano adottarla indipendentemente dall'implementazione dell'autore. Questa sezione registra brevemente quei termini: sono governance anziché metodo, e nulla nei §3–§8 dipende da essi. Quattro impegni li fissano.

Una specifica aperta e irrevocabile. La specifica del Δ-CSI è pubblicata sotto licenza aperta (CC BY), come definizione aperta e irrevocabile anziché come permesso ritirabile in seguito. Chiunque può calcolare e usare l'indice agli stessi termini pubblici.

Parità arm's-length. L'implementazione di riferimento è *una* implementazione conforme, non lo standard. Calcola il Δ-CSI sotto la stessa definizione pubblica disponibile a qualsiasi altro implementatore, e non gode di alcuna variante privilegiata. L'autorità dello standard poggia sulla definizione, non sull'implementazione.

Un'identità citabile propria. La specifica ha un'identità — un nome, un documento, un DOI — distinta dalla documentazione di prodotto di qualsiasi implementazione. Ciò che altri citano quando citano il Δ-CSI è così lo standard, non il manuale di un fornitore.

Questo paper è quella definizione citabile, e prevale sulle descrizioni interne precedenti dell'indice sulla questione della definizione stessa. I documenti interni all'implementazione restano autorevoli sul comportamento dell'implementazione, e subordinati a questa specifica sulla definizione.

Un impegno alla custodia indipendente. La custodia dello standard, a questo stadio, spetta all'autore come persona fisica. È sufficiente per una definizione la cui priorità e paternità sono stabilite dalla pubblicazione.

L'impegno qui registrato è che il nome e la specifica siano conferiti a un custode indipendente — strutturalmente separato, nel controllo e nel finanziamento, da qualsiasi singolo implementatore — nel punto in cui l'adozione esterna renda necessaria quella separazione. Lo scopo è enunciato con chiarezza: un indice del giudizio indipendente non può essere governato in modo credibile, nel lungo periodo, da una parte che lucra sui punteggi. Porre lo standard a termini aperti e irrevocabili ora è ciò che tiene aperta quell'opzione futura, anziché precluderla.

Questi impegni sono governance, non matematica, e non cambiano nulla nel §4. Sono inclusi perché l'alternativa minerebbe, a livello di custodia, l'indipendenza che l'indice esiste per misurare a livello di decisioni: uno standard di misura la cui fonte canonica è un documento interno di prodotto, rivedibile a discrezione del fornitore.

SEZIONE 10

Conclusione

Questo paper ha definito il Δ-CSI — Cognitive Sovereignty Index, delta — come l'intensità della sfida strutturata che un contraddittore cognitivo applica a una singola decisione. E ha specificato come quell'intensità sia calcolata: deterministicamente, da cinque componenti limitati e scomponibili, sotto regole anti-gaming e di confrontabilità fissate in una specifica versionata.

Ha tracciato il confine del costrutto in quattro negazioni: l'indice non è qualità della decisione, non è probabilità di successo, non è un giudizio sul decisore, e non è se la decisione sia cambiata. Sono le quattro confusioni che il nome e la scala dell'indice invitano attivamente a commettere, e una misura dell'intensità della sfida che le permettesse non misurerebbe nulla di utilizzabile. Ed è stato esplicito, per tutto il testo, sulla linea tra ciò che è definito, ciò che è ingegnerizzato e ciò che resta da mostrare.

Ciò che esiste al momento della scrittura è una definizione, un metodo riproducibile e una disciplina. Un numero che due parti qualsiasi possono calcolare in modo identico dalla stessa sfida e attribuire alle sue parti. Difese che, in assenza di una sfida reale, fanno fallire il sistema in modo visibile anziché adulare in silenzio. Un registro di debolezze fatte emergere dallo stesso metodo avversariale che l'indice applica alle decisioni che valuta.

Ciò che non esiste ancora è la validità empirica — evidenza che una data intensità della sfida corrisponda a qualcosa sulle decisioni scorate — e il paper non l'ha né affermata né oscurata nella sua assenza.

La validazione è esposta come programma pre-registrato: la calibrazione del ledger pertenant, e il pilot di forecasting sigillato del companion, in cui l'indice entra come provenienza e covariata, mai come grandezza a punteggio. Il ledger di esiti vuoto è riportato per quello che è.

La postura non è incidentale rispetto all'oggetto. Un costrutto costruito per misurare quanto onestamente una decisione affronti ciò che preferirebbe non esaminare non può essere introdotto da un documento che risparmia a sé stesso lo stesso trattamento.

Il primo e più esigente test del Δ-CSI è quello che questo paper ha cercato di superare: enunciare una misura con precisione, difenderla dove può essere difesa, contraddirla dove non può. E lasciare che il numero stia solo per ciò che effettivamente misura, che è l'intensità della sfida, e non la qualità della decisione.

APPENDICE

Δ-CSI Specification v1.0

La specifica normativa del Δ-CSI è redatta in inglese come versione di riferimento unica: in caso di divergenza interpretativa prevale il testo inglese qui riportato. È la definizione a cui rimandano il §4 e il §9 del paper. I riferimenti interni usano la forma A.N (es. A.9.1); i «§N» senza prefisso rimandano al corpo del paper. È congelata; ogni modifica richiede un cambio di versione.

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.

RIFERIMENTI

Riferimenti

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* [Rapporto di sintesi, giugno 2026]. Roma. *(Il Cognitive Sovereignty Index organizzativo è introdotto in questo lavoro; editore/sede da finalizzare per la versione di riferimento.)*

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. [Ristampato 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. [Opera originale pubblicata come *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. Pubblicato a 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. Pubblicato in *Advances in Neural Information Processing Systems 36* (NeurIPS 2023), Datasets and Benchmarks Track. https://arxiv.org/abs/2306.05685