FRANCESCO SAVERIO CANEPA · WORKING PAPER
FRAMEWORK / THEORY PAPER

Sovranità cognitiva: proteggere la qualità delle decisioni delle classi dirigenti nell'era dell'intelligenza artificiale

Companion teorico di Calibrated Dissent (osf.io/5qc8t) — pre-registrato su OSF il 25 giugno 2026

di

Francesco Saverio Canepa

VERSIONE 1.0 · 29 GIUGNO 2026

INDICE

Indice

Sintesi

Le classi dirigenti che operano con assistenza AI intensiva fronteggiano un paradosso strutturale: l'abbondanza di analisi può aggravare — non correggere — i meccanismi attraverso cui il giudizio si erode. I sistemi AI ottimizzati per la soddisfazione dell'utente tendono a compiacere, a gonfiare la confidenza e, nel tempo, a omologare il ragionamento tra organizzazioni che condividono gli stessi strumenti. Questo paper propone il costrutto di **sovranità cognitiva** — la capacità collettiva di mantenere giudizio indipendente, calibrato e falsificabile in queste condizioni — e ne specifica le dimensioni operative, i meccanismi di erosione e le condizioni di osservabilità.

Il costrutto integra quattro linee di letteratura — sycophancy nei modelli RLHF (Sharma et al., 2023), over-reliance (Parasuraman & Manzey, 2010), deskilling (Bainbridge, 1983) e monocultura algoritmica (Kleinberg & Raghavan, 2021) — in quattro dimensioni: indipendenza di giudizio, resistenza all'omologazione, calibrazione dell'incertezza e tracciabilità dell'onere della prova. Il meccanismo di protezione proposto è il **contraddittore cognitivo**, definito da quattro principi invarianti — avversarialità strutturale, fail-closed sulla compiacenza, schema di output fisso, separazione provenienza/scoring — che lo distinguono da un validatore sofisticato; da essi il paper argomenta un quadro di governance per board (sfida obbligatoria, registro deliberativo prospettico, separazione dei ruoli, calibrazione nel tempo). Il Δ-CSI — proxy calcolato dell'intensità della sfida applicata in una sessione (Canepa, 2026) — è reinterpretato come aggancio parziale del costrutto, non come sua misura complessiva né predittore di correttezza. L'Osservatorio — registro pubblico append-only di N = 10 previsioni M&A sigillate prima dell'esito — ne illustra l'istanziazione su decisioni reali.

Il paper **non** afferma che il contraddittore migliori le decisioni, che la sovranità cognitiva predica la correttezza, né che le dimensioni siano già operativizzate o validate empiricamente: il costrutto è una proposta teorica da operativizzare. Il caso N = 10 è non-evidenza per qualsiasi affermazione di efficacia, situandosi al di sotto della soglia di potenza necessaria per qualunque test confermativo (Canepa, 2026, §6.5); la generalizzabilità oltre i domini ad alto rischio illustrati è una questione empirica aperta. Questo paper è il companion teorico di *Calibrated Dissent* (Canepa, 2026), pre-registrato su OSF (osf.io/5qc8t), dove è documentato il materiale tecnico condiviso.

SEZIONE 1

Introduzione

Le classi dirigenti operano oggi in una condizione inedita: accesso a sistemi AI capaci di sintetizzare evidenze, elaborare scenari e restituire analisi più rapide e articolate di qualsiasi staff precedente.

Il paradosso è che questo ampliamento dell'analisi può aggravare — non correggere — il problema fondamentale di ogni decisione consequenziale: il modello del mondo su cui poggia viene raramente messo alla prova prima di essere tradotto in azione; più spesso, viene silenziosamente confermato. La tendenza a cercare evidenze che corroborino le ipotesi preferite — documentata da Kahneman (2011) come uno dei bias più robusti del ragionamento sotto incertezza — non è un semplice limite individuale correggibile con più informazioni: è una disposizione sistemica che si aggrava quando le informazioni sono abbondanti, preselezionate e confezionate per compiacere. Gli assistenti AI ottimizzati per la soddisfazione dell'utente realizzano esattamente questa condizione — un'inclinazione che la ricerca recente identifica come bias strutturale indotto dall'addestramento (Sharma et al., 2023): setacciano le evidenze alla ricerca di sostegno alla tesi prevalente e consegnano la validazione che un decisore sotto pressione desidera ricevere. La compiacenza non è un difetto caratteriale dello strumento; è un esito progettuale.

Per chi esercita funzioni dirigenziali — consiglieri di amministrazione, comitati esecutivi, investitori istituzionali, regolatori — questa dinamica assume contorni particolarmente critici. Le decisioni ad alto rischio non si prendono in assenza di informazioni, ma *in presenza* di informazioni abbondanti, selezionate e interpretate attraverso il filtro delle premesse che si è già disposti ad accettare. Quando lo strumento AI incaricato di assistere quella valutazione tende strutturalmente a rinforzare la lettura dominante, il risultato non è un'elaborazione aumentata ma un'erosione sistematica della capacità di giudizio indipendente, mascherata da rigore analitico. Vi si aggiunge il rischio di una monocultura cognitiva: quando organizzazioni diverse impiegano i medesimi sistemi AI, la convergenza degli strumenti può produrre convergenza del giudizio (Kleinberg & Raghavan, 2021), erodendo la diversità di prospettiva da cui i processi deliberativi sani dipendono.

Il cambio di paradigma necessario non è tecnico ma architettonico: da *oracolo* a *contraddittore cognitivo*. La genealogia di questa distinzione è consolidata. Schwenk (1990) ha esaminato evidenze sperimentali e sul campo, documentando come il devil's advocacy e l'inquiry dialettico — metodi che impongono il contro-argomento esplicito — tendano a produrre, in media, piani strategici di qualità superiore rispetto ai contesti

in cui il dissenso è soltanto consentito; quella letteratura precede interamente i modelli linguistici. Cai, Arawjo e Glassman (2024) hanno poi formalizzato lo spazio di progettazione per agenti AI la cui funzione di utilità è esplicitamente oppositiva — la cosiddetta *antagonistic AI* — aprendo la domanda se un LLM possa operativizzare il devil's advocacy su scala, in modo robusto e non soltanto performativo. Come renderla affidabile è uno dei temi centrali di questo paper (§4).

La categoria concettuale che questo paper introduce è la **sovranità cognitiva**: la capacità di una classe dirigente di mantenere giudizio indipendente, calibrato e falsificabile in un contesto in cui i sistemi AI tendono a omologare e compiacere. Il costrutto, sviluppato in §3, sistematizza le condizioni di erosione del giudizio dirigenziale — compiacenza strumentale, over-reliance, deskilling progressivo e monocultura cognitiva — e propone le dimensioni attraverso cui tale capacità può essere riconosciuta, monitorata e protetta. Il Δ-CSI, già introdotto come proxy dell'intensità della sfida in *Calibrated Dissent* (Canepa, 2026), è qui reinterpretato come uno strumento di aggancio del costrutto più ampio — non l'unico, e non una misura di correttezza decisionale.

Questo paper si posiziona come companion teorico di *Calibrated Dissent* (Canepa, 2026), pre-registrato su OSF il 25 giugno 2026: quello è un paper di metodi e protocollo che affronta *come* misurare il dissenso calibrato senza senno di poi; questo concettualizza *perché* la qualità del giudizio delle classi dirigenti sia strutturalmente esposta e quali principi di design e di governance possano proteggerla. Il materiale tecnico condiviso — pipeline a tre passate, schema di output a sette campi, definizione operativa del Δ-CSI — è citato da *Calibrated Dissent*, non replicato qui.

Il paper articola tre contributi. Il primo è il costrutto di sovranità cognitiva: definizione, dimensioni operative, antecedenti, modalità di erosione e condizioni di osservabilità (§3). Il secondo è una generalizzazione dei principi di design del contraddittore cognitivo come meccanismo di protezione, riusabili e non proprietari (§4). Il terzo è un quadro di governance per board: chi sfida, quando, come la decisione è registrata e auditata (§6). L'arco segue il percorso *problema → meccanismo → adozione*.

SEZIONE 2

Il problema: come l'IA degrada le decisioni dirigenziali

Quattro meccanismi distinti ma strutturalmente connessi erodono la qualità del giudizio delle classi dirigenti in presenza di sistemi AI: la compiacenza generata dall'addestramento, l'eccessiva fiducia nell'output automatico, il deterioramento progressivo delle competenze di giudizio esperto e l'omologazione del ragionamento collettivo. Nessuno è esclusivamente un problema dell'AI generativa; ciascuno ha radici in letterature più antiche — ergonomia cognitiva, organizational behavior, teoria delle decisioni — che i modelli linguistici moderni proiettano su scala e velocità senza precedenti. Prima di proporre un costrutto unificante (§3) è necessario esaminare ciascun meccanismo con precisione, distinguere le evidenze fondate dalle ipotesi plausibili e identificare la lacuna concettuale aperta.

2.1 — SYCOPHANCY NEI LLM: LA COMPIACENZA COME MODALITÀ DI GUASTO

I modelli linguistici di grandi dimensioni addestrati con reinforcement learning from human feedback (RLHF) mostrano una propensione documentata a modificare le proprie valutazioni per accordarsi con le preferenze — reali o percepite — dell'interlocutore. Uno studio sistematico su modelli di diverse dimensioni e gradi di fine-tuning RLHF (Sharma et al., 2023) ha identificato questo fenomeno come *sycophancy*: la tendenza a spostare le risposte in direzione delle posizioni espresse dall'utente, incluse posizioni deliberatamente erronee introdotte nella conversazione, in modo positivamente associato al grado di ottimizzazione per la soddisfazione dell'utente. Il meccanismo proposto è che il segnale di reward usato nell'addestramento incentiva le risposte gradite all'annotatore umano, e il modello generalizza quella preferenza verso forme di accordo e validazione indipendentemente dalla correttezza della posizione sostenuta dall'utente.

Questo risultato va interpretato con precisione: non afferma che i modelli RLHF siano inutili né che ogni risposta sia un'eco della posizione dell'utente, ma che — in modo sistematico e misurabile attraverso protocolli controllati — l'addestramento per la soddisfazione introduce un bias verso la compiacenza. È un esito progettuale, non un'imperfezione contingente: l'ottimizzazione per un proxy (l'approvazione dell'annotatore) diverge dall'obiettivo epistemico del decisore, il giudizio accurato e indipendente. Per i contesti deliberativi ad alto rischio, in cui il giudizio indipendente è il bene più critico da proteggere, questa divergenza non è trascurabile.

La sycophancy individuale non esaurisce il problema. I sistemi multi-agente — più LLM che interagiscono tra loro, spesso proposti come antidoto alle limitazioni dei mod-

elli singoli — non eliminano necessariamente la deriva compiacente ma possono trasferirla al livello dell'interazione tra agenti. Un'analisi delle modalità di fallimento nel debate multi-agente (Wynn et al., 2025) ha documentato come le discussioni tra LLM possano collassare in falso disaccordo o convergenza prematura anziché produrre una sfida genuina; uno studio controllato su compiti di ragionamento logico (Wu et al., 2025) ha riscontrato che il debate produce benefici limitati e dipendenti dalla struttura del compito. La moltiplicazione degli agenti non è dunque di per sé una soluzione alla compiacenza strutturale: il mandato avversariale deve essere incorporato nell'architettura del sistema, non soltanto nell'interfaccia.

2.2 — AUTOMATION BIAS E OVER-RELIANCE: LA CESSIONE DEL CONTROLLO CRITICO

La sycophancy sarebbe meno grave se i decisori vi resistessero sistematicamente, integrando l'output AI come un'informazione tra molte, valutata criticamente alla luce del proprio giudizio esperto. L'evidenza sull'interazione uomo-automazione suggerisce che l'ipotesi è ottimistica.

Parasuraman e Manzey (2010) hanno sviluppato un modello integrato della compiacenza nell'uso dell'automazione, sintetizzando decenni di ricerca su sistemi semi-automatici in aviazione, medicina, controllo industriale e supervisione di processi complessi. Il loro quadro distingue la *compiacenza dell'automazione* — la riduzione della vigilanza quando il sistema è percepito come affidabile — dall'*automation bias* (qui e nel seguito anche *over-reliance*): la tendenza sistematica a trattare la raccomandazione automatizzata come punto focale del ragionamento, attribuendole un peso che eccede quello giustificato dalla qualità effettiva del sistema. Entrambi i fenomeni si manifestano indipendentemente dall'accuratezza reale del sistema: la percezione di affidabilità, una volta stabilita, resiste all'aggiornamento anche di fronte ad anomalie rilevabili.

Le implicazioni per i processi decisionali dirigenziali sono dirette. Quando un sistema AI produce un'analisi articolata, strutturata e apparentemente fondata su evidenze, il decisore è soggetto alle stesse dinamiche che Parasuraman e Manzey descrivono per gli operatori di sistemi automatici: trattare l'output come punto di partenza privilegiato, ridurre la ricerca di informazioni contrarie e abbassare la vigilanza critica nella misura in cui il sistema appare affidabile. In contesti ad alto rischio — dove gli errori dell'automazione sono rari ma consequenziali — questa riduzione dell'attenzione è la più pericolosa: genera fallimenti concentrati nei casi in cui il sistema sbaglia con maggiore sicurezza apparente.

2.3 — DESKILLING: IL DETERIORAMENTO SILENZIOSO DELL'EXPERTISE

Un terzo meccanismo opera su orizzonti più lunghi ed è perciò più difficile da osservare in tempo reale: l'uso prolungato di sistemi automatici erode le competenze di giudizio indipendente che quei sistemi dovrebbero supportare. La formulazione classica resta quella di Bainbridge (1983) nel saggio «Ironies of Automation», ancora rilevante a oltre quattro decenni.

Bainbridge identificò un paradosso strutturale nell'ingegneria dei sistemi automatici: i progettisti assegnano all'operatore il compito di intervenire quando l'automazione fallisce — nei casi più critici e atipici — ma l'automazione stessa, riducendo la frequenza di esercizio delle competenze necessarie, atrofizza le capacità richieste per quell'intervento. Più il sistema è affidabile nell'uso normale, meno l'operatore pratica le abilità richieste nelle condizioni di guasto, e meno sarà in grado di intervenire quando servono. L'ironia è duplice: il progresso dell'automazione non semplifica il ruolo cognitivo dell'operatore ma lo trasforma, richiedendo un picco di competenza esattamente quando la pratica di quella competenza è stata minimizzata dall'automazione stessa.

La trasposizione a contesti di leadership non è metaforica. Un dirigente che demanda sistematicamente agli assistenti AI la sintesi di contesti competitivi, la valutazione dei rischi e la costruzione di scenari alternativi non smette di decidere, ma riduce la frequenza con cui costruisce autonomamente quei modelli del mondo. Nel breve periodo l'impatto è invisibile; nel medio, la capacità di riconoscere i limiti dell'analisi AI, di interrogarne le premesse con competenza disciplinare e di produrre giudizi indipendenti quando lo strumento fallisce — tipicamente nei casi fuori dalla distribuzione su cui è stato addestrato — si deteriora silenziosamente. Il deskilling non cancella il giudizio: lo svuota progressivamente di contenuto critico.

2.4 — OMOLOGAZIONE COGNITIVA: IL RISCHIO DELLA MONOCULTURA ALGORITMICA

I tre meccanismi descritti operano principalmente a livello individuale o organizzativo. Un quarto, di natura sistemica, emerge considerando l'insieme dei decisori di un settore o di un'istituzione: la convergenza degli strumenti AI può produrre una convergenza dei ragionamenti le cui conseguenze collettive non sono riducibili alla somma degli effetti individuali.

Kleinberg e Raghavan (2021) hanno analizzato formalmente le implicazioni della *monocultura algoritmica* — la condizione in cui molte organizzazioni adottano lo stesso algoritmo di supporto alle decisioni — in modelli di selezione e allocazione. La loro analisi mostra che gli esiti sistemici della monocultura possono differire da quelli di un'ad-

ozione diversificata, anche quando ogni singolo adottante ne beneficia individualmente. Il meccanismo è la correlazione degli errori: quando molti agenti usano lo stesso modello predittivo, i loro errori sistematici cessano di essere indipendenti e diventano covarianti, erodendo il beneficio di diversificazione che in condizioni di eterogeneità ammortizza le valutazioni erronee dei singoli.

Trasportato all'AI applicata alle decisioni dirigenziali, il risultato ha valenza immediata. Quando comitati di investimento, consigli di amministrazione e analisti di organizzazioni diverse si rivolgono agli stessi sistemi AI, non cedono soltanto individualmente in autonomia di giudizio: contribuiscono a una struttura di correlazioni che riduce la diversità sistemica del ragionamento disponibile. Le valutazioni divergenti — che in mercati e processi deliberativi sani portano alla superficie informazioni che la lettura dominante trascura — diventano più rare, non per accordo esplicito ma per convergenza degli strumenti. L'omologazione cognitiva non è una minaccia solo futura: è già, nei settori a elevata penetrazione AI, una condizione strutturale in atto.

2.5 — LA GENEALOGIA DELLA SFIDA STRUTTURATA: DAL DEVIL'S ADVOCACY ALL'IA AVVERSARIALE

I quattro meccanismi non sono privi di antidoti concettuali. La letteratura sulla sfida strutturata offre una genealogia consolidata, che precede interamente i modelli linguistici moderni e va esaminata per collocare il contributo di questo paper nella continuità delle idee.

Schwenk (1990) ha esaminato l'evidenza sperimentale e sul campo sul *devil's advocacy* e sulla *dialectical inquiry* come tecniche prescrittive per il miglioramento delle decisioni strategiche. La sua rassegna documenta che i gruppi in cui un agente ha il mandato istituzionale di argomentare contro la proposta prevalente producono, in media, valutazioni di qualità superiore rispetto a quelli in cui il dissenso è soltanto consentito ma non strutturalmente richiesto. Il risultato centrale non è che il disaccordo sia utile in quanto tale, ma che la sua istituzionalizzazione — l'incorporazione nell'architettura del processo piuttosto che nel carattere dei partecipanti — appaia essere una condizione necessaria per la sua efficacia.

Quella tradizione ha trovato, quattro decenni più tardi, un'eco nel lavoro di Cai, Arawjo e Glassman (2024), che hanno formalizzato lo spazio di progettazione per sistemi AI la cui funzione di utilità è esplicitamente oppositiva — la cosiddetta *antagonistic AI* — articolando le tensioni tra sistemi orientati alla soddisfazione dell'utente e sistemi orientati a un'utilità epistemica più esigente, che include la sfida attiva al ragionamento del decisore. L'articolazione di questo spazio è concettualmente necessaria prima di poter discutere un sistema AI che operativizzi il devil's advocacy in modo affidabile e non meramente performativo.

La lacuna che questa genealogia lascia aperta è precisa. Né la letteratura sulla sfida strutturata né quella sull'antagonistic AI offrono un costrutto teorico che colleghi i

quattro meccanismi di erosione — compiacenza, over-reliance, deskilling, monocultura cognitiva — a una nozione unitaria di *capacità di giudizio indipendente* riconoscibile, monitorabile e proteggibile a livello di governance. Il devil's advocacy è una tecnica procedurale, l'antagonistic AI uno spazio di progettazione, la monocultura algoritmica un rischio sistemico: ciascuna elaborazione resta circoscritta al proprio dominio, senza un quadro integrato che descriva la condizione cognitiva della classe dirigente nell'era dell'AI e ne specifichi le condizioni di tutela. È questo costrutto unificante — la *sovranità cognitiva* — che §3 introduce e sviluppa.

SEZIONE 3

Il costrutto: sovranità cognitiva

La lacuna identificata in §2.5 è concettuale: i quattro meccanismi di erosione — compiacenza, over-reliance, deskilling, monocultura — sono stati studiati come fenomeni separati, ciascuno nella propria letteratura, senza un costrutto che li colleghi a un'unica capacità minacciata. Questa sezione lo propone come proposta teorica da operativizzare, non come grandezza già misurata o validata: ne seguono la definizione, l'articolazione delle dimensioni, l'analisi dell'erosione e le condizioni di osservabilità.

3.1 — DEFINIZIONE FORMALE

Definiamo la **sovranità cognitiva** di una classe dirigente come *la capacità collettiva di mantenere un giudizio indipendente, calibrato e falsificabile sulle decisioni consequenziali, a fronte di sistemi di intelligenza artificiale che tendono strutturalmente a compiacere l'utente e a omologare il ragionamento tra decisori*. La definizione contiene tre qualificatori non negoziabili. Il giudizio è *indipendente* quando la sua direzione non è determinata dal pull conformativo dell'output AI né dal consenso degli strumenti adottati dai pari. È *calibrato* quando le stime di confidenza che lo accompagnano tracciano le frequenze empiriche degli esiti, nel senso della letteratura sulla previsione accurata (Tetlock & Gardner, 2015). È *falsificabile* quando è formulato in modo da specificare quali evidenze lo confuterebbero e su chi gravi l'onere di produrle.

Questa nozione va distinta da due concetti contigui. Non coincide con lo *scetticismo*: lo scetticismo è la disposizione a sospendere l'assenso e, elevato a regola, produce paralisi o rifiuto indiscriminato dell'evidenza, mentre la sovranità cognitiva richiede l'assenso calibrato — credere nella misura giustificata dalle evidenze, né più né meno. Non coincide neppure con l'*autonomia* generica: si può decidere senza vincoli esterni mantenendo però un giudizio mal calibrato o conformista. La sovranità cognitiva è più stringente — è autonomia *di un giudizio che resta indipendente, calibrato e falsificabile sotto la pressione specifica esercitata dai sistemi AI*: una proprietà del giudizio sotto carico avversariale, non del decisore in astratto.

Due precisazioni completano la definizione. La sovranità cognitiva è un attributo di una *classe dirigente* — un consiglio, un comitato, una funzione deliberativa — non del singolo: è una proprietà del processo collettivo con cui un'organizzazione forma e mette alla prova i propri giudizi, e si erode o si protegge a quel livello. Ed è *graduale*, non binaria: non si possiede o si perde, ma si tiene in misura maggiore o minore lungo le dimensioni che seguono.

3.2 — DIMENSIONI

Proponiamo quattro dimensioni del costrutto. Ciascuna ha una definizione operativa — la grandezza che, in linea di principio, la renderebbe osservabile — e risponde a uno specifico meccanismo di erosione di §2. Sono proposte come articolazione del costrutto da operativizzare e validare, non come scale già misurate.

(1) Indipendenza di giudizio. È la resistenza del giudizio al pull conformativo dell'output AI: la misura in cui la posizione finale del decisore diverge, quando le evidenze lo giustificano, da quella che l'assistente AI ha validato o suggerito. Operativamente si manifesta nello scarto tra il giudizio formato prima e dopo l'esposizione all'output AI, condizionato alla qualità delle evidenze: un decisore con elevata indipendenza aggiorna in direzione dell'evidenza, non dell'output. Risponde direttamente alla compiacenza documentata nei modelli RLHF (Sharma et al., 2023): la sycophancy è il vettore che spinge il giudizio verso la posizione gradita, e l'indipendenza ne è la contromisura a livello di capacità.

(2) Resistenza all'omologazione. È la divergenza preservata rispetto al consenso di modello: la misura in cui il giudizio di un'organizzazione resta distinto da quello che altre, alimentate dagli stessi sistemi AI, formerebbero dalle medesime evidenze. Operativamente si osserva nella correlazione tra i giudizi di decisori diversi che condividono lo stesso strumento — bassa correlazione (a parità di evidenze) indica resistenza preservata, alta correlazione omologazione. Risponde al rischio di monocultura algoritmica (Kleinberg & Raghavan, 2021): dove la convergenza degli strumenti rende covarianti gli errori sistematici, questa è la capacità che mantiene la diversità di giudizio da cui i processi deliberativi sani dipendono.

(3) Calibrazione dell'incertezza. È la proprietà per cui le stime di confidenza tracciano le frequenze empiriche degli esiti: tra le valutazioni espresse con confidenza del 70%, circa il 70% dovrebbe risultare corretto a posteriori. Operativamente è la grandezza misurata dalle curve di calibrazione e dagli score della letteratura sulla previsione (Tetlock & Gardner, 2015). Risponde all'effetto congiunto di due meccanismi di §2: la compiacenza gonfia la confidenza fornendo validazione non meritata (Sharma et al., 2023), mentre l'over-reliance induce ad attribuire all'output automatico una fiducia che eccede la sua affidabilità effettiva (Parasuraman & Manzey, 2010). La calibrazione è la dimensione che entrambi questi vettori erodono dal lato della stima di incertezza.

(4) Tracciabilità dell'onere della prova. È la proprietà per cui, in una decisione, è esplicito e registrato chi deve dimostrare cosa: quale parte sostiene la tesi, quale ha il mandato di confutarla, quali asserzioni restano non provate e quali evidenze le confermerebbero o falsificherebbero. Operativamente si osserva nella presenza, nel processo deliberativo, di un tracciato verificabile che attribuisce l'onere probatorio e ne registra l'assolvimento o il mancato assolvimento. Risponde all'over-reliance (Parasuraman &

ale non interrogato del ragionamento, l'onere della prova si dissolve silenziosamente — nessuno è più tenuto a dimostrare alcunché, perché lo strumento «ha già analizzato» — e la capacità stessa di formulare quale prova servirebbe si atrofizza con il disuso.

3.3 — ANTECEDENTI ED EROSIONE

Le quattro dimensioni non si indeboliscono in modo indipendente: i meccanismi di §2 sono antecedenti causali che le erodono per vie distinte e spesso simultanee. La compiacenza strumentale (§2.1) attacca indipendenza (1) e calibrazione (3); l'over-reliance (§2.2), dissolvendo l'attribuzione dell'onere probatorio e inducendo una fiducia mal tarata, attacca tracciabilità (4) e calibrazione (3); il deskilling (§2.3), erodendo la pratica del giudizio autonomo, attacca indipendenza (1) e tracciabilità (4); l'omologazione cognitiva (§2.4) opera in modo sistemico sulla resistenza all'omologazione (2), rendendo covarianti i giudizi di organizzazioni che condividono gli strumenti.

Due proprietà di questa erosione meritano enfasi, perché spiegano perché la sovranità cognitiva richieda protezione deliberata anziché potersi assumere spontaneamente stabile. La prima è la *silenziosità*: nessuno dei quattro meccanismi produce un segnale di allarme mentre agisce. La compiacenza si presenta come accordo articolato, l'over-reliance come efficienza, il deskilling è invisibile nel breve periodo (§2.3), la monocultura emerge solo all'aggregazione di molte decisioni. La tendenza a cercare conferma alle ipotesi preferite (Kahneman, 2011) fa sì che il decisore non percepisca l'erosione come perdita ma come conferma del proprio giudizio. La seconda è la *covarianza*: i meccanismi non si sommano ma si rinforzano. La compiacenza alimenta l'over-reliance — un output che valida è un output di cui ci si fida — che accelera il deskilling, il quale a sua volta riduce la capacità di resistere alla compiacenza. È questa dinamica di rinforzo reciproco, non un singolo meccanismo isolato, a rendere la condizione cognitiva della classe dirigente strutturalmente esposta.

3.4 — OSSERVABILITÀ E MISURA

Un costrutto è scientificamente utile nella misura in cui specifica le condizioni alle quali diventa osservabile. La sovranità cognitiva, per come è definita in §3.1, non è osservabile direttamente: è una capacità, non un evento. Diventa osservabile solo attraverso le sue manifestazioni lungo le quattro dimensioni, e a tre condizioni. La prima è una *traccia decisionale*: le dimensioni (1), (3) e (4) richiedono che il giudizio prima dell'esposizione all'output AI, la confidenza dichiarata e l'attribuzione dell'onere probatorio siano registrati al momento della decisione, non ricostruiti a posteriori — un registro deliberativo prospettico è la precondizione di ogni osservazione. La seconda è la *risoluzione degli esiti*: la calibrazione (3) non è valutabile finché un numero sufficiente di giudizi non sia stato confrontato con gli esiti. La terza è il *confronto tra decisori*: la resistenza all'omologazione (2) è una grandezza relativa, osservabile solo confrontando i

giudizi di più organizzazioni a parità di evidenze, non misurabile su un singolo decisore isolato.

In questo quadro, il **Δ-CSI** (Cognitive Sovereignty Index, delta) introdotto in *Calibrated Dissent* (Canepa, 2026) è *uno* degli strumenti possibili di aggancio del costrutto — non l'unico, e non una misura della sovranità cognitiva nel suo complesso. Per la definizione operativa data in quel lavoro (Canepa, 2026, §2.3), è un **proxy dell'intensità della sfida applicata** in una sessione — quanto forte sia stata la pressione esercitata sul ragionamento — **e non una misura della correttezza della decisione**; trattarlo come segnale di accuratezza è un errore di categoria. Tra le quattro dimensioni aggancia soprattutto la tracciabilità dell'onere della prova (4), di cui registra sessione per sessione l'attività di confutazione; non misura invece l'indipendenza (1), la resistenza all'omologazione (2) né la calibrazione (3), che richiedono rispettivamente il confronto pre/post-esposizione, il confronto tra decisori e la risoluzione degli esiti. La sovranità cognitiva non è dunque riducibile al Δ-CSI: costruire strumenti per le altre tre dimensioni — e validarli empiricamente — resta un programma di ricerca aperto, non un risultato di questo paper, che propone il costrutto e ne specifica le condizioni di osservabilità senza affermare di averlo misurato.

SEZIONE 4

Il meccanismo: principi di design del contraddittore cognitivo

La sezione precedente ha identificato le quattro dimensioni del costrutto e i rispettivi antecedenti di erosione. Riconoscere il costrutto non basta: occorre un meccanismo che operi attivamente per proteggerlo. È il **contraddittore cognitivo** — un sistema progettato non per assistere il ragionamento del decisore, ma per sfidarlo. La sfida è la funzione di utilità del sistema, non un sotto-prodotto occasionale di un'analisi orientata alla sintesi.

Il termine *contraddittore cognitivo* non designa un prodotto specifico né un'architettura brevettata: designa una classe di sistemi che condividono un insieme di principi progettuali necessari — quattro, come argomentiamo di seguito — senza i quali un sistema non può dirsi contraddittore in senso sostanziale, qualunque etichetta adotti. Non sono configurazioni opzionali, ma invarianti architettoniche che distinguono il contraddittore dal validatore sofisticato. Come la letteratura sul devil's advocacy documenta, l'efficacia della sfida strutturata dipende dalla sua istituzionalizzazione — dall'incorporazione nell'architettura del processo, non nel carattere dei partecipanti (Schwenk, 1990); lo stesso vale per un sistema AI che ambisce a operativizzare quel mandato su scala.

I quattro principi sono riusabili da chiunque progetti sistemi AI con mandato avversariale. Ciascuno risponde a uno o più meccanismi di erosione di §2 e protegge dimensioni specifiche del costrutto di §3. Un'implementazione di riferimento che li soddisfa tutti e quattro — CounterBrain — è discussa al termine della sezione con l'unico scopo di mostrare che i principi non sono puramente teorici ma realizzabili in un contesto operativo.

4.1 — AVVERSARIALITÀ STRUTTURALE

Il primo principio stabilisce che il mandato del sistema sia sfidare la decisione, mai validarla. È più stringente di quanto sembri: non richiede che il sistema produca *anche* una sfida accanto all'analisi, ma che la sfida *sia* l'analisi. La validazione non è un obiettivo secondario da bilanciare con l'opposizione: è una modalità incompatibile con il ruolo del contraddittore.

Cai, Arawjo e Glassman (2024) hanno formalizzato lo spazio di progettazione per l'AI antagonista, articolando il gradiente tra sistemi orientati alla soddisfazione dell'utente e sistemi il cui obiettivo è esplicitamente oppositivo. L'avversarialità strutturale colloca il contraddittore sull'estremo oppositivo di quel gradiente senza configurazioni intermedie: non esiste una modalità «contraddittore moderato» o «sfida ammorbidita». Non

è un limite di ergonomia ma un principio di integrità progettuale: un sistema la cui intensità avversariale sia configurabile ritornerebbe, per inerzia organizzativa o per volontà del committente, verso l'estremo validante — esattamente l'ottimizzazione che l'addestramento per la soddisfazione dell'utente incentiva strutturalmente (Sharma et al., 2023).

L'evidenza sperimentale rafforza questa posizione. Schwenk (1990) ha documentato che i gruppi in cui il devil's advocacy è *richiesto* producono decisioni strategiche di qualità superiore rispetto a quelli in cui è semplicemente *consentito*, perché la permissività consente di omettere la sfida proprio quando è più necessaria — quando il gruppo è già convinto della propria tesi. Il mandato strutturale rimuove quella via di fuga. Per un sistema AI il corollario è diretto: se la modalità avversariale è configurabile, sarà disabilitata precisamente dai decisori più esposti alla compiacenza.

La dimensione che questo principio protegge in via primaria è l'**indipendenza di giudizio** (§3.2, dimensione 1): un sistema il cui mandato sia validare non può, per definizione, preservare il carattere non-conformativo del giudizio. La sycophancy comprime la distanza tra il giudizio del decisore e la posizione validata dallo strumento (Sharma et al., 2023); l'avversarialità strutturale vi oppone una forza opposta, sistematica e non modulabile. Il medesimo mandato tutela anche la **resistenza all'omologazione** (dimensione 2): costringendo a divergere dalla lettura dominante anziché convergere verso il consenso del modello, preserva la distanza di giudizio tra l'organizzazione e i pari che condividono gli stessi strumenti.

4.2 — FAIL-CLOSED SULLA COMPIACENZA

Il secondo principio specifica la modalità di guasto del sistema. Se l'output di una passata non soddisfa la soglia di sfida obbligatoria — se cioè il sistema sta per emettere un'analisi che non sfida genuinamente la tesi — deve sollevare un errore esplicito, non emettere silenziosamente l'output non sfidante. La compiacenza è una modalità di guasto, non degradata: l'equivalente funzionale di un errore di tipo II, dove il mancato rifiuto di un'ipotesi falsa è un fallimento del test, non un risultato neutro.

Una sfida vuota o minimale che passi per valida — in assenza di un gate esplicito — è indistinguibile dall'output di un contraddittore robusto per l'osservatore che non può ispezionare la pipeline. Il decisore è allora falsamente rassicurato dalla *forma* dell'output (ha l'aspetto di una sfida strutturata) mentre il contenuto è privo di forza avversariale: una modalità di guasto più grave dell'assenza di sfida, perché rimuove la percezione del rischio senza eliminarne la sostanza.

La ricerca sui sistemi multi-agente documenta fenomeni corrispondenti: debate tra LLM che collassano in falso disaccordo o convergenza prematura (Wynn et al., 2025) e compiti di ragionamento in cui il debate produce benefici limitati e dipendenti dalla struttura del compito (Wu et al., 2025). La moltiplicazione degli agenti non è dunque di per sé una protezione contro il fail-open: serve un gate esplicito che verifichi, dopo

ogni passata avversariale, che la sfida sia non vuota su ciascuna delle dimensioni obbligatorie, con un retry dedicato e, al secondo fallimento, un errore di motore che interrompe la pipeline senza emettere output parziali.

Il principio fail-closed protegge in via primaria la **tracciabilità dell'onere della prova** (§3.2, dimensione 4). Il meccanismo corrispondente è l'over-reliance (Parasuraman & Manzey, 2010): quando un output AI appare strutturalmente completo, il decisore tende a trattarlo come punto focale, riducendo l'attenzione nella verifica della forza avversariale. Il fail-closed assicura che questa fiducia non venga accordata a output che non la meritano; la protezione è progettuale, non dipendente dalla vigilanza del decisore.

4.3 — SCHEMA DI OUTPUT FISSO

Il terzo principio stabilisce che l'output del contraddittore sia conforme a uno schema con campi obbligatori che nessuna specializzazione verticale o configurazione cliente può rimuovere, rinominare o rendere facoltativi. La struttura fissa non è una preferenza stilistica: è la traduzione architettonica dell'invarianza della sfida, la sua esteriorizzazione in un contratto verificabile.

Calibrated Dissent (Canepa, 2026, §2.2) specifica uno schema a sette campi — costruzione, proprietà e motivazioni documentate campo per campo in quella sede — versionato come contratto JSON con `additionalProperties: false`. Non lo riproduciamo qui: ciò che interessa è la *funzione di principio* dello schema fisso, indipendente dai valori specifici adottati per i sette campi.

Quella funzione è triplice. Primo, ogni campo obbligatorio corrisponde a una manifestazione specifica dell'erosione del costrutto: le assunzioni implicite mettono in visibilità le premesse che la sycophancy lascerebbe silenziosamente confermate; i test di falsificazione traducono l'onere della prova in verifiche empiriche azionabili; la confidenza calibrata, con i gap di conoscenza dichiarati, contrasta il gonfiamento della fiducia indotto dall'over-reliance (Parasuraman & Manzey, 2010). Secondo, la fissità dello schema consente comparazione longitudinale e controllo di qualità: se un campo può essere omesso, viene meno la garanzia che la sfida di una sessione sia comparabile con le precedenti, e l'accumulo di dati per la calibrazione resta privo di base. Terzo, il carattere non rinominabile dei campi previene la deriva lessicale — il rischio che una configurazione alternativa reintroduca gli stessi campi con un vocabolario che ne affievolisce il mandato avversariale (ad esempio rinominando «test di falsificazione» come «osservazioni complementari»).

Lo schema fisso protegge trasversalmente tutte e quattro le dimensioni, con enfasi particolare sulla **calibrazione dell'incertezza** (dimensione 3) — la presenza obbligatoria di una confidenza calibrata e di gap di conoscenza espliciti è la condizione necessaria affinché le stime siano confrontabili con gli esiti realizzati — e sulla **tracciabilità**

dell'onere della prova (dimensione 4), per la struttura dei campi di falsificazione e delle domande sull'onere probatorio come elementi distinti e obbligatori.

4.4 — SEPARAZIONE PROVENIENZA/SCORING

Il quarto principio riguarda una dipendenza circolare che i sistemi di analisi con retrieval possono introdurre in assenza di una progettazione esplicita: il materiale recuperato come provenienza non deve alterare retroattivamente le quantità sottoposte a giudizio.

Il rischio è il seguente. Un contraddittore dotato di strumenti di ricerca e recupero può reperire, nel corso dell'analisi, evidenze che modificano la formulazione della sfida. Se quelle stesse evidenze influenzano simultaneamente il contenuto dell'output e la misurazione dell'intensità — in particolare il Δ-CSI, proxy dell'intensità della sfida applicata (Canepa, 2026, §2.3) — si introduce una circolarità in cui provenienza e scoring si co-determinano: valori elevati del Δ-CSI possono allora essere un artefatto della disponibilità di fonti abbondanti, non di un'intensità avversariale genuinamente applicata al ragionamento. L'indice cessa di misurare la proprietà che si intende monitorare.

La separazione si realizza architettonicamente isolando la fase di retrieval da quella di scoring: le fonti vengono recuperate e registrate come provenienza, ma le componenti del Δ-CSI sono calcolate deterministicamente dalla struttura logica della sfida — assunzioni, test di falsificazione, domande sull'onere della prova, divergenza scenaristica — non dal volume o dalla qualità del materiale recuperato. Il conteggio delle fonti contribuisce all'indice come componente separata con peso proprio, ma non retroagisce sulle altre né altera la struttura logica della sfida già prodotta. È la condizione affinché il Δ-CSI resti interpretabile come proxy dell'intensità della sfida, invariante rispetto al contesto informativo, e non come misura della correttezza della decisione — distinzione che il paper tratta come invariante interpretativo (§3.4; Canepa, 2026, §2.3).

Il principio di separazione protegge in via primaria la **calibrazione dell'incertezza** (dimensione 3): la validità empirica di un indice di intensità della sfida richiede che esso misuri una proprietà del ragionamento, non del retrieval. Un Δ-CSI confuso con l'abbondanza delle fonti non può essere calibrato nel senso della letteratura sulla previsione — come corrispondenza tra stime di confidenza dichiarate e frequenze empiriche degli esiti (Tetlock & Gardner, 2015).

COUNTERBRAIN — IMPLEMENTAZIONE DI RIFERIMENTO

I quattro principi di questa sezione — avversarialità strutturale, fail-closed sulla compiacenza, schema di output fisso, separazione provenienza/scoring — sono condizioni necessarie per qualsiasi sistema che intenda operare come contraddittore cognitivo; non sono proprietari né vincolati a una specifica architettura o organizzazione. CounterBrain, il sistema di cui *Calibrated Dissent* (Canepa, 2026) riporta il protocollo di misurazione, è l'implementazione di riferimento che mostra la realizzabilità di questi principi in un contesto operativo: avversarialità codificata come invariante non configurabile; fail-closed come gate esplicito con retry e interruzione della pipeline in assenza di sfida valida; schema a sette campi versionato come contratto JSON con `additionalProperties: false`; separazione provenienza/scoring garantita architettonicamente (Canepa, 2026, §2.4). Sistemi alternativi che soddisfino gli stessi quattro principi rispetterebbero ugualmente il mandato del contraddittore. La menzione serve a un unico scopo epistemico: collocare i principi nel dominio dell'operazionalizzabile, non dell'ideale teorico.

SEZIONE 5

Caso illustrativo: l'Osservatorio (N = 10)

> AVVERTENZA METODOLOGICA
>
> **ILLUSTRA, NON PROVA.** Il materiale presentato in questa sezione illustra che il dispositivo delineato in §3 e §4 è stato effettivamente istanziato su decisioni reali. Non costituisce in alcun modo un test dell'efficacia del contraddittore cognitivo, non consente di valutare se le previsioni siano calibrate, e non offre evidenza confermativa di alcuna ipotesi del programma di ricerca. Il test confermativo è riservato a una Fase 2 potenziata, pre-registrata nel paper companion (Canepa, 2026); questa sezione è il resoconto dell'avvio del dispositivo, non del suo verdetto.

5.1 — IL DISPOSITIVO: REGISTRO PUBBLICO, APPEND-ONLY, ANCORATO TRAMITE HASH

Il costrutto di sovranità cognitiva (§3) richiede un registro deliberativo prospettico: il giudizio deve essere fissato prima che qualsiasi esito sia osservabile, in una forma che preclude la revisione retroattiva — senza di che la calibrazione misura solo un'illusione costruita a posteriori. Il meccanismo del contraddittore (§4) richiede parimenti che la sfida sia strutturale, che l'output rispetti uno schema fisso e non rinominabile e che la provenienza non alteri retroattivamente le quantità sottoposte a scoring (§4.4). La domanda operativa è se un dispositivo che soddisfi simultaneamente queste condizioni sia realizzabile fuori dal laboratorio, su decisioni reali e consequenziali, con garanzie di non-manipolabilità verificabili da terzi.

L'**Osservatorio** è la risposta concreta a quella domanda: un registro pubblico append-only in cui ogni scheda di previsione — prodotta dal processo contraddittore — è pubblicata prima che qualunque esito sia noto. L'integrità è garantita da un doppio ancoraggio. Il primo è un hash SHA-256 del contenuto al momento del sigillo: qualsiasi modifica successiva produce un hash diverso, rendendo la revisione immediatamente rilevabile. Il secondo è un'ancora OpenTimestamps (OTS), che incorpora l'hash in una transazione sulla blockchain di Bitcoin, la cui immutabilità dipende dalla potenza computazionale distribuita dell'intera rete, non dalla fiducia in un singolo operatore. La coppia sigillo-SHA256 + ancora-OTS realizza il requisito di tracciabilità dell'onere della prova (§3.2, dimensione 4): qualunque revisore esterno può verificare in modo indipendente che il testo pubblicato corrisponda all'hash originale e che quell'hash fosse registrato sulla blockchain prima di un dato istante.

Il protocollo di sigillo e risoluzione — le regole deterministiche per stabilire quando un esito si è realizzato, le fonti canoniche, la doppia codifica cieca raccomandata — è documentato in Canepa (2026, §4–§5). Questa sezione aggiunge solo la constatazione che il protocollo è stato effettivamente avviato su decisioni reali.

5.2 — IL PILOT: N = 10 PREVISIONI SIGILLATE IL 20 GIUGNO 2026

Il **20 giugno 2026** (T_0) sono state sigillate nell'Osservatorio **dieci previsioni su altrettante operazioni di fusione e acquisizione europee ad alto rischio**. Ciascuna scheda pubblica registra: la distribuzione di scenario a tre vie prodotta dalla Passata 1 del contraddittore (lettura dominante), la tesi principale esplicitata, una stima P(H1) — deal-break o rinegoziazione al ribasso ≥ 15% nell'orizzonte 6–12 mesi — e, per il sottoinsieme pre-dichiarato di quattro operazioni con acquirente quotato che consolida il target, una stima P(H2) per l'orizzonte 18–36 mesi. Il payload tecnico integrale del motore — schema a sette campi, Δ-CSI e bundle hash che congela modello, prompt e Pack — è conservato come provenienza non sottoposta a scoring (§4.4) e non compare nelle schede pubbliche al momento del sigillo.

La pre-registrazione della coorte è pubblica: **osf.io/5qc8t**, DOI 10.17605/OSF.IO/5QC8T (Canepa, 2026). Il deposito OSF è avvenuto il 25 giugno 2026 senza embargo: il materiale pre-specificato è liberamente accessibile alla pubblicazione di questo paper. L'Osservatorio è consultabile all'indirizzo counterbrain.com/registro, dove schede sigillate e trail di risoluzione progressivo sono disponibili senza autenticazione.

Le finestre di risoluzione si estendono da giugno 2027 (H1, 12 mesi) a giugno 2029 (H2, 36 mesi). Al momento della stesura di questo paper nessun esito è ancora osservabile sull'intero campione.

5.3 — PERIMETRO EPISTEMICO: COSA IL CASO ILLUSTRA E COSA NON AFFERMA

ILLUSTRA, NON PROVA. Il caso risponde a una sola domanda: i principi di design del contraddittore cognitivo (§4) sono realizzabili in un contesto operativo su decisioni reali? La risposta che il caso consente è affermativa nel senso dell'esistenza e dell'operatività: l'avversarialità strutturale è stata applicata a previsioni consequenziali, non a stimoli ipotetici di laboratorio; lo schema a sette campi è stato prodotto e versionato; la pipeline fail-closed ha operato su un insieme definito di casi; la separazione provenienza/scoring è implementata come invariante architettonico; il registro append-only con ancoraggio OTS esiste ed è verificabile da chiunque. Sul versante del costrutto (§3), la condizione di osservabilità prospettica — la registrazione del giudizio prima dell'esito — è soddisfatta, e la tracciabilità dell'onere della prova (dimensione 4) è registrata per ciascuna scheda.

Questo è il perimetro epistemico esatto. Oltre quel confine, il caso non afferma nulla. Non afferma che il contraddittore sia efficace. Non afferma che le previsioni siano calibrate. Non afferma che i principi di design producano decisioni di qualità superiore. Non afferma che il Δ-CSI predica la correttezza delle decisioni sfidate: il Δ-CSI,

ove citato nelle schede di provenienza, è esclusivamente un proxy dell'intensità della sfida applicata in quella sessione (§4.4; Canepa, 2026, §2.3), cioè misura con quanta forza il processo ha premuto sulle dimensioni del costrutto, non se la decisione sfidata abbia buone probabilità di successo. Trattarlo come segnale di accuratezza è un errore di categoria (§3.4). N = 10 è strutturalmente al di sotto della soglia di potenza necessaria per qualunque test confermativo — fissata a N ≥ 20 per orizzonte in Canepa (2026, §6.5) — e gli esiti sono in larga misura ancora futuri.

Il test confermativo — la valutazione della calibrazione (H-cal) e della divergenza-che-paga (H-pay) rispetto alle baseline pre-registrate — è riservato a una **Fase 2 potenziata**, con una coorte più ampia sigillata in avanti, le cui condizioni di analisi sono pre-registrate in Canepa (2026, §6.5). Quella fase produrrà evidenza confermativa o, se le ipotesi non reggono, la loro falsificazione. L'Osservatorio è il dispositivo che rende possibile un test onesto senza senno di poi; non ne anticipa il verdetto.

SEZIONE 6

Governance: proteggere la sovranità cognitiva a livello di board

Il problema (§2) ha definito i meccanismi attraverso cui l'AI degrada il giudizio dirigenziale; il meccanismo (§4) ha specificato i principi che un contraddittore cognitivo deve rispettare per operare in modo genuinamente avversariale. La terza fase dell'arco — l'adozione — affronta una domanda distinta: come una classe dirigente può istituzionalizzare la funzione di sfida nella propria architettura deliberativa? La risposta non è tecnica ma organizzativa: richiede protocolli espliciti, attribuzioni di ruolo, registri decisionali e meccanismi di calibrazione che operino strutturalmente, non per vigilanza episodica dei singoli.

6.1 — INQUADRAMENTO: NORMARNE L'USO, NON VIETARLO

La tentazione di affrontare i rischi di §2 limitando l'AI nelle decisioni ad alto rischio è comprensibile ma mal indirizzata. La compiacenza dei modelli (Sharma et al., 2023), l'over-reliance (Parasuraman & Manzey, 2010) e il deskilling (Bainbridge, 1983) non dipendono dall'intensità dell'uso ma dalla sua architettura: lo stesso sistema usato senza sfida strutturale produce erosione, con un contraddittore istituzionalizzato produce pressione epistemica. Il problema non è l'AI nelle decisioni consequenziali — è l'AI senza contraddittore.

La lezione della letteratura sulla sfida strutturata è che la variabile determinante nella qualità delle decisioni strategiche non è la competenza dei partecipanti né la quantità di informazioni, bensì l'architettura del processo (Schwenk, 1990). La governance del contraddittore deve ereditarla: il mandato di sfida va codificato in quell'architettura, non delegato al carattere o alla vigilanza dei singoli.

6.2 — QUATTRO PRINCIPI DI GOVERNANCE

I seguenti quattro principi traducono la logica progettuale di §4 al livello organizzativo. Sono formulati come pratiche di prudenza istituzionale per organi deliberativi che operano in condizioni di elevata assistenza AI — non come derivazione da obblighi normativi specifici già codificati, ma come risposta razionale alle dinamiche di erosione documentate.

(i) Sfida obbligatoria al punto decisionale. Ogni decisione che superi una soglia di materialità pre-definita deve attivare un mandato di sfida prima che l'organo agisca. L'obbligatorietà è il nucleo del principio: una sfida che può essere omessa tende a esserlo nei casi di maggiore pressione — esattamente quelli in cui sarebbe più necessaria

(Schwenk, 1990). La soglia va definita ex ante da una funzione non interessata alla decisione, su criteri di valore della posta, reversibilità e portata strategica; la reversibilità è la più critica, perché una decisione non reversibile a costo ragionevole entro un orizzonte predefinito è candidata alla sfida indipendentemente dal valore monetario.

(ii) Registrazione e audit della decisione e della sfida. Una decisione consequenziale deve generare un tracciato deliberativo che registri: la questione formulata prima del ricorso all'AI; l'output del processo di sfida, con assunzioni critiche, test di falsificazione proposti e confidenza con i gap di conoscenza dichiarati; la deliberazione dell'organo e i fattori che l'hanno determinata; la decisione assunta. L'onere di identificare assunzioni critiche e test di falsificazione è in capo alla funzione di sfida; l'organo motiva esplicitamente perché la decisione regge nonostante la sfida. Il tracciato è prodotto al momento della decisione, non a posteriori: senza registrazione prospettica, la razionalizzazione post-esito — rinforzata dalle stesse tendenze compiacenti dei modelli AI (Sharma et al., 2023) — rende la calibrazione empiricamente irraggiungibile (Tetlock & Gardner, 2015).

(iii) Separazione dei ruoli: chi decide ≠ chi sfida. La funzione di sfida non deve essere affidata alla stessa persona o unità che ha preparato l'analisi originale. È il parallelo organizzativo dell'avversarialità strutturale (§4.1): se chi sfida ha un interesse nell'adeguatezza dell'analisi che ha prodotto, la sfida tende verso la validazione per le stesse ragioni motivazionali che inducono i modelli RLHF verso la sycophancy (Sharma et al., 2023). La separazione si realizza con rotazione delle responsabilità, nomina esterna per categorie specifiche di decisioni o — nelle implementazioni AI-assistite — isolamento architettonico della funzione di sfida; ciò che esclude è l'assetto in cui chi produce l'analisi giudica anche se essa è stata adeguatamente messa alla prova.

(iv) Calibrazione verificata nel tempo. Un organo che applichi la sfida strutturata in modo sistematico accumula un registro di decisioni, stime di confidenza e, progressivamente, esiti. A intervalli regolari — almeno annuali, o al raggiungimento di un numero sufficiente di decisioni risolte — quel registro va esaminato: le stime erano calibrate nel senso di Tetlock e Gardner (2015)? I test di falsificazione erano azionabili o meramente formali? Emergono fallimenti sistematici in categorie specifiche? Non è una valutazione della performance individuale, ma un audit di calibrazione del processo, progettato per rilevare la deriva lenta — il deskilling descritto da Bainbridge (1983) — prima che renda inassessabile la qualità delle decisioni.

6.3 — IL REGISTRO DELIBERATIVO COME PATTERN ORGANIZZATIVO

I principi (ii) e (iv) convergono verso un'architettura comune: il **registro deliberativo**. L'esigenza è semplice ma stringente — il tracciato decisionale deve precedere l'esito, non essere retroattivamente modificabile ed essere accessibile alla funzione di calibrazione. È una specifica funzionale, non una tecnologia: qualsiasi implementazione che

assicuri registrazione prospettica, struttura append-only e verificabilità indipendente soddisfa il requisito.

Questo pattern è talvolta denominato *Sovereign Decision Ledger* — un termine che designa il pattern organizzativo, non un sistema specifico. Il principio è che giudizio, sfida e attribuzione dell'onere della prova siano fissati prima che qualsiasi esito sia osservabile, in una forma che preclude la revisione silenziosa, senza cui la tracciabilità dell'onere della prova (§3.2) resta nominale. L'Osservatorio di §5 ne è un'istanziazione operativa, con architettura documentata in Canepa (2026). Il pattern è indipendente dalla tecnologia e scalabile: applicarlo richiede disciplina procedurale, non infrastruttura specializzata.

La condizione minima perché il registro abbia valore epistemico è che il tracciato sia prodotto prima che il decisore conosca l'esito, anche parziale: qualsiasi retroazione dell'esito sull'attribuzione dell'onere della prova invalida la calibrazione e trasforma il registro in razionalizzazione documentata.

6.4 — CONNESSIONE AI DOVERI FIDUCIARI: UNA QUESTIONE DI PRUDENZA EPISTEMICA

I principi di governance di questa sezione non derivano da un quadro normativo specificamente codificato — non si afferma che una prescrizione di legge li richieda. Sono però coerenti con la logica dei doveri fiduciari che governano gli organi deliberativi nella maggior parte degli ordinamenti. Il dovere di diligenza richiede che i decisori agiscano su base informata, in buona fede, con la cura di una persona ordinariamente prudente. Operare in condizioni documentate di over-reliance (Parasuraman & Manzey, 2010) — in cui l'output AI diventa sistematicamente il punto focale del ragionamento, con un peso che eccede l'affidabilità effettiva del sistema — senza un meccanismo strutturale di sfida pone interrogativi sulla compatibilità con quello standard di prudenza, sul piano dei suoi stessi presupposti cognitivi e non per effetto di una prescrizione esterna.

In questa prospettiva, la sfida obbligatoria, il tracciato deliberativo e la calibrazione periodica non sono ornamenti procedurali: sono le condizioni strutturali che rendono accessibile lo standard di deliberazione informata che la prudenza istituzionale richiede. Il deskilling (Bainbridge, 1983) e l'omologazione cognitiva (Kleinberg & Raghavan, 2021) operano su orizzonti che i singoli organi non percepiscono in tempo reale: fronteggiarli richiede meccanismi istituzionali, non vigilanza individuale episodica.

L'efficacia di questi principi applicati specificamente alla deliberazione AI-assistita a livello di board non è stata testata in modo controllato; l'evidenza disponibile proviene dalla letteratura organizzativa (Schwenk, 1990) e dalla ricerca sulla calibrazione (Tetlock & Gardner, 2015), la cui trasferibilità al contesto specifico richiede cautela. I principi offrono una traduzione coerente della logica di §4 al livello istituzionale; il registro

deliberativo prodotto dal principio (iv) è il meccanismo attraverso cui un'organizzazione può raccogliere nel tempo evidenze sull'efficacia del proprio dispositivo di sfida.

SEZIONE 7

Limiti e cosa questo paper non afferma

Questo paper propone un costrutto teorico e i principi di design del meccanismo corrispondente; ne enuncia qui i confini epistemici in modo esplicito, distinguendo ciò che afferma da ciò che potrebbe essere erroneamente letto come sottinteso.

Sull'efficacia del contraddittore cognitivo. Questo paper non afferma che il contraddittore cognitivo migliori la qualità delle decisioni. I quattro principi di design di §4 — avversarialità strutturale, fail-closed sulla compiacenza, schema di output fisso, separazione provenienza/scoring — sono condizioni necessarie per un sistema che aspiri a operativizzare il devil's advocacy in modo non meramente performativo; non sono condizioni sufficienti a garantire decisioni di qualità superiore, né un mandato di efficacia causale dimostrato. L'efficacia è una domanda empirica che questo paper non risolve e non pretende di risolvere.

Sulla sovranità cognitiva come predittore di correttezza. Questo paper non afferma che la sovranità cognitiva predica la correttezza delle decisioni. Le quattro dimensioni del costrutto (§3.2) descrivono le condizioni di qualità del processo deliberativo, non la bontà dei suoi esiti. Una classe dirigente può preservare intatta la propria sovranità cognitiva e assumere comunque decisioni errate; può eroderla sistematicamente e, per circostanze favorevoli, ottenere risultati positivi. La sovranità cognitiva è una proprietà del processo, non un predittore dell'esito. Analogamente, il Δ-CSI — proxy dell'intensità della sfida applicata in una sessione (Canepa, 2026, §2.3) — non predice la correttezza: un valore elevato segnala che il processo ha premuto con forza su molte dimensioni, non che la decisione abbia buone probabilità di successo. Trattarlo come segnale di accuratezza è un errore di categoria (§3.4).

Sullo statuto empirico del costrutto. Il costrutto di sovranità cognitiva è una proposta teorica, non una grandezza misurata né validata empiricamente. Le quattro dimensioni sono articolate con definizioni operative che ne specificano le condizioni di osservabilità; non sono scale già costruite, testate e psicometricamente validate. Il programma di operativizzazione — la traduzione delle definizioni operative in strumenti di misura, il test della struttura fattoriale, la stima della validità convergente e discriminante rispetto a costrutti contigui — è un'agenda di ricerca futura, dichiarata esplicitamente in §8. Il Δ-CSI aggancia una sola dimensione (la tracciabilità dell'onere della prova, §3.4) ed è per definizione un proxy dell'intensità della sfida, non una misura del costrutto nel suo complesso.

Sul caso illustrativo N = 10. Il pilot dell'Osservatorio presentato in §5 è non-evidenza per qualsiasi affermazione di efficacia del contraddittore. Come esplicitato in §5.3, il

caso rende disponibile evidenza di istanziazione operativa: i principi di design di §4 sono stati applicati a decisioni reali e la condizione di tracciabilità prospettica è soddisfatta; non si ricava alcuna evidenza che il contraddittore produca previsioni calibrate, che si distingua da una baseline di tasso-base, o che incida sulla qualità delle decisioni sfidate. N = 10 si situa strutturalmente al di sotto della soglia di potenza necessaria per qualunque test confermativo, fissata a N ≥ 20 per orizzonte in Canepa (2026, §6.5). Gli esiti sono in larga misura ancora futuri; il test confermativo è riservato alla Fase 2 potenziata pre-registrata in quel lavoro.

Sulla generalizzabilità. Il costrutto è sviluppato con riferimento primario a contesti di decisione ad alto rischio e bassa reversibilità — M&A, investimenti istituzionali, deliberazioni strategiche di board. Non si afferma che sia trasferibile senza riesame a domini a bassa posta, ad alta frequenza decisionale o con una struttura di incentivi radicalmente diversa. Che i meccanismi di erosione (§2) si manifestino nelle stesse forme e che i principi di design e di governance li contrastino con la stessa efficacia in altri domini è una questione empirica aperta.

Sull'interpretazione della letteratura. I meccanismi di erosione di §2 e i principi di §4 poggiano su letteratura con statuto scientifico indipendente — Schwenk (1990), Bainbridge (1983), Sharma et al. (2023), Kleinberg e Raghavan (2021). L'interpretazione che ne viene fornita, e soprattutto la sua applicazione alla condizione delle classi dirigenti AI-assistite, è un contributo teorico che attende validazione; il paper non afferma che sia l'unica lettura possibile né che la letteratura citata esaurisca il panorama rilevante.

Conclusione e agenda di ricerca

Questo paper ha argomentato lungo un arco in tre mosse: problema, meccanismo, adozione. Il problema (§2) è strutturale: i sistemi AI ottimizzati per la soddisfazione dell'utente erodono il giudizio dirigenziale attraverso quattro meccanismi distinti e rinforzanti — compiacenza strumentale, over-reliance, deskilling progressivo e monocultura cognitiva — che operano silenziosamente e si aggravano sotto elevata assistenza AI. La lacuna aperta è l'assenza di un costrutto unificante che li colleghi a una capacità collettiva — la sovranità del giudizio — riconoscibile, monitorabile e proteggibile a livello di governance.

Il meccanismo (§3–§4) risponde a quella lacuna. La sovranità cognitiva è la capacità collettiva di mantenere giudizio indipendente, calibrato e falsificabile nelle condizioni di pressione descritte, e le sue quattro dimensioni ne specificano l'osservabilità. Il contraddittore cognitivo è il meccanismo di protezione proposto: un sistema la cui funzione di utilità è la sfida, non la validazione, vincolato a quattro principi invarianti affinché la sfida sia sostanziale e non performativa. L'adozione (§6) traduce quella logica al livello organizzativo: sfida obbligatoria al punto decisionale, registrazione prospettica e auditabile della deliberazione, separazione dei ruoli tra chi produce l'analisi

e chi la sfida, calibrazione verificata nel tempo. Il registro deliberativo — append-only e non retroattivamente modificabile — è la precondizione che rende accessibile la calibrazione empirica e con essa il miglioramento del processo (Tetlock & Gardner, 2015).

Agenda di ricerca. Tre priorità emergono come passi necessari perché la proposta teorica possa diventare un contributo empirico.

La prima è l'operativizzazione e validazione delle dimensioni del costrutto. Le definizioni operative di §3.2 specificano cosa ciascuna dimensione misurerebbe; la traduzione in strumenti psicometrici testabili — scale, protocolli di osservazione, procedure di codifica — è un programma non avviato. La struttura fattoriale, la validità convergente e quella discriminante rispetto a costrutti contigui restano da stabilire empiricamente.

La seconda è il test confermativo della Fase 2. Come documentato in *Calibrated Dissent* (Canepa, 2026, §6.5), il protocollo di sigillo-e-risoluzione dell'Osservatorio è progettato per consentire, su una coorte più ampia e sigillata in avanti, un test pre-registrato delle ipotesi di calibrazione (H-cal) e di divergenza-che-paga (H-pay). Quella fase produrrà evidenza sulla qualità previsionale del contraddittore — o, se le ipotesi non reggono, la loro falsificazione — ed è il passo necessario per trasformare il claim di efficacia da proposta teorica in affermazione empiricamente fondata o rigettata.

La terza è l'estensione oltre l'M&A. I principi del costrutto e del meccanismo si applicano in linea di principio a qualsiasi decisione ad alto rischio con bassa reversibilità: politica pubblica, investimento istituzionale, governance sanitaria in condizioni di incertezza. Verificare se i meccanismi di erosione (Bainbridge, 1983; Schwenk, 1990) si manifestano nelle stesse forme — e se gli stessi principi li contrastano con la stessa efficacia — richiede ricerca indipendente in quei domini.

Conflitto di interessi

L'autore è fondatore di CounterBrain / Coopera S.a.s. CounterBrain è citato in §4 esclusivamente come implementazione di riferimento che illustra la realizzabilità dei quattro principi progettuali in un contesto operativo; la menzione non implica alcuna raccomandazione commerciale né una valutazione comparativa dell'adeguatezza di quella implementazione rispetto ad alternative. I principi di design del contraddittore cognitivo enunciati in §4 sono presentati come non proprietari e riusabili da chiunque progetti sistemi AI con mandato avversariale. Nessuna fonte di finanziamento esterna ha influenzato il contenuto di questo paper.

Riferimenti

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

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

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

4 Kahneman, D. (2011). *Thinking, Fast and Slow*. New York: Farrar, Straus and Giroux.

5 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

6 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

7 Schwenk, C. R. (1990). Effects of Devil's Advocacy and Dialectical Inquiry on Decision Making. *Organizational Behavior and Human Decision Processes*, 47(1), 161–176. https://doi.org/10.1016/0749-5978(90)90037-P

8 Sharma, M., Tong, M., Korbak, T., Duvenaud, D., Askell, A., Bowman, S. R., Cheng, N., Durmus, E., Hatfield-Dodds, Z., Johnston, S. R., Kravec, S., Maxwell, T., McCandlish, S., Ndousse, K., Rausch, O., Schiefer, N., Yan, D., Zhang, M., & Perez, E. (2023). *Towards Understanding Sycophancy in Language Models*. arXiv:2310.13548. https://arxiv.org/abs/2310.13548

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

10 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. https://arxiv.org/abs/2511.07784

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