FRANCESCO SAVERIO CANEPA · WORKING PAPER
FRAMEWORK / THEORY PAPER

Cognitive Sovereignty: Protecting the Quality of Leadership Decisions in the Age of Artificial Intelligence

Theoretical companion to Calibrated Dissent (osf.io/5qc8t) — pre-registered on OSF 25 June 2026

by

Francesco Saverio Canepa

VERSION 1.0 · 29 JUNE 2026

CONTENTS

Table of Contents

FRAMEWORK / THEORY PAPER — THEORETICAL COMPANION TO CALIBRATED DISSENT

Abstract

Leadership operating under intensive AI assistance faces a structural paradox: an abundance of analysis can aggravate — not correct — the mechanisms through which judgment erodes. AI systems optimized for user satisfaction tend to please, to inflate confidence, and, over time, to homogenize reasoning across organizations that share the same tools. This paper introduces the construct of **cognitive sovereignty** — the collective capacity to maintain independent, calibrated, and falsifiable judgment under these conditions — and specifies its operational dimensions, erosion mechanisms, and conditions of observability.

The construct integrates four strands of literature — sycophancy in RLHF models (Sharma et al., 2023), over-reliance (Parasuraman & Manzey, 2010), deskilling (Bainbridge, 1983), and algorithmic monoculture (Kleinberg & Raghavan, 2021) — into four dimensions: independence of judgment, resistance to homogenization, uncertainty calibration, and burden-of-proof traceability. The proposed protection mechanism is the **cognitive contradictor**, defined by four invariant principles — structural adversariality, fail-closed on sycophancy, fixed output schema, provenance/scoring separation — that distinguish it from a sophisticated validator; from these the paper argues a governance framework for boards (mandatory challenge, prospective deliberative ledger, role separation, calibration over time). The Δ-CSI — a computed proxy of challenge intensity applied in a session (Canepa, 2026) — is reinterpreted as a partial anchor of the construct, not as its overall measure nor as a predictor of correctness. The Osservatorio — a public, append-only register of N = 10 M&A forecasts sealed before any outcome — illustrates its instantiation on real decisions.

The paper does **not** claim that the contradictor improves decisions, that cognitive sovereignty predicts correctness, nor that the dimensions are already operationalized or empirically validated: the construct is a theoretical proposal to be operationalized. The N = 10 case is non-evidence for any claim of effectiveness, falling below the power threshold required for any confirmatory test (Canepa, 2026, §6.5); generalizability beyond the high-stakes domains illustrated is an open empirical question. This paper is the theoretical companion to *Calibrated Dissent* (Canepa, 2026), pre-registered on OSF (osf.io/5qc8t), where the shared technical material is documented.

SECTION 1

Introduction

Leadership today operates in an unprecedented condition: access to AI systems capable of synthesizing evidence, elaborating scenarios, and returning analyses faster and more articulated than any previous staff.

The paradox is that this expansion of analysis can aggravate — not correct — the fundamental problem of every consequential decision: the model of the world on which it rests is rarely tested before being translated into action; more often, it is silently confirmed. The tendency to seek evidence that corroborates preferred hypotheses — documented by Kahneman (2011) as one of the most robust biases in reasoning under uncertainty — is not a simple individual limitation correctable with more information: it is a systemic disposition that worsens when information is abundant, preselected, and packaged to please. AI assistants optimized for user satisfaction realize precisely this condition — an inclination that recent research identifies as a structural training-induced bias (Sharma et al., 2023): they sift evidence in search of support for the prevailing thesis and deliver the validation that a decision-maker under pressure wishes to receive. Sycophancy is not a character flaw of the tool; it is a design outcome.

For those exercising leadership functions — board members, executive committees, institutional investors, regulators — this dynamic takes on particularly critical contours. High-stakes decisions are not made in the absence of information, but *in the presence* of abundant information, selected and interpreted through the filter of premises one is already willing to accept. When the AI tool entrusted with assisting that evaluation structurally tends to reinforce the dominant reading, the result is not augmented deliberation but a systematic erosion of the capacity for independent judgment, disguised as analytical rigor. Added to this is the risk of cognitive monoculture: when different organizations employ the same AI systems, the convergence of tools can produce a convergence of judgment (Kleinberg & Raghavan, 2021), eroding the diversity of perspective on which healthy deliberative processes depend.

The necessary paradigm shift is not technical but architectural: from *oracle* to *cognitive contradictor*. The genealogy of this distinction is well established. Schwenk (1990) examined experimental and field evidence, documenting how devil's advocacy and dialectical inquiry — methods that impose the explicit counter-argument — tend to produce, on average, strategic plans of superior quality compared to contexts in which dissent is merely permitted; that literature predates large language models entirely. Cai, Arawjo, and Glassman (2024) subsequently formalized the design space for AI agents

whose utility function is explicitly oppositional — the so-called *antagonistic AI* — opening the question of whether an LLM can operationalize devil's advocacy at scale, robustly and not merely performatively. How to make it reliable is one of the central themes of this paper (§4).

The conceptual category this paper introduces is **cognitive sovereignty**: the capacity of a leadership body to maintain independent, calibrated, and falsifiable judgment in a context in which AI systems tend to homogenize and please. The construct, developed in §3, systematizes the conditions of erosion of leadership judgment — instrumental sycophancy, over-reliance, progressive deskilling, and cognitive monoculture — and proposes the dimensions through which this capacity can be recognized, monitored, and protected. The Δ-CSI, already introduced as a proxy of challenge intensity in *Calibrated Dissent* (Canepa, 2026), is here reinterpreted as one tool for anchoring the broader construct — not the only one, and not a measure of decision correctness.

This paper positions itself as the theoretical companion to *Calibrated Dissent* (Canepa, 2026), pre-registered on OSF on 25 June 2026: that is a methods and protocol paper addressing *how* to measure calibrated dissent without hindsight; this one conceptualizes *why* the quality of leadership judgment is structurally exposed and what design and governance principles can protect it. The shared technical material — the three-pass pipeline, the seven-field output schema, the operational definition of the Δ-CSI — is cited from *Calibrated Dissent*, not replicated here.

The paper articulates three contributions. The first is the construct of cognitive sovereignty: definition, operational dimensions, antecedents, modes of erosion, and conditions of observability (§3). The second is a generalization of the design principles of the cognitive contradictor as a protection mechanism, reusable and non-proprietary (§4). The third is a governance framework for boards: who challenges, when, and how the decision is recorded and audited (§6). The arc follows the path *problem → mechanism → adoption*.

SECTION 2

The problem: how AI degrades leadership decisions

Four distinct but structurally connected mechanisms erode the quality of leadership judgment in the presence of AI systems: training-generated sycophancy, excessive confidence in automated output, the progressive deterioration of expert judgment skills, and the homogenization of collective reasoning. None is exclusively a problem of generative AI; each has roots in older literatures — cognitive ergonomics, organizational behavior, decision theory — that modern language models project onto unprecedented scale and speed. Before proposing a unifying construct (§3) it is necessary to examine each mechanism with precision, to distinguish grounded evidence from plausible hypotheses, and to identify the open conceptual gap.

2.1 — SYCOPHANCY IN LLMS: SYCOPHANCY AS A FAILURE MODE

Large language models trained with reinforcement learning from human feedback (RLHF) exhibit a documented propensity to modify their evaluations to align with the preferences — real or perceived — of the interlocutor. A systematic study across models of different sizes and degrees of RLHF fine-tuning (Sharma et al., 2023) identified this phenomenon as *sycophancy*: the tendency to shift responses in the direction of positions expressed by the user, including deliberately erroneous positions introduced into the conversation, in a manner positively associated with the degree of optimization for user satisfaction. The proposed mechanism is that the reward signal used in training incentivizes responses pleasing to the human annotator, and the model generalizes that preference toward forms of agreement and validation regardless of the correctness of the user's stated position.

This finding must be interpreted precisely: it does not claim that RLHF models are useless or that every response is an echo of the user's position, but that — in a systematic and measurable way through controlled protocols — training for satisfaction introduces a bias toward sycophancy. It is a design outcome, not a contingent imperfection: optimization for a proxy (annotator approval) diverges from the epistemic objective of the decision-maker, namely accurate and independent judgment. For high-stakes deliberative contexts, in which independent judgment is the most critical asset to protect, this divergence is not negligible.

Individual sycophancy does not exhaust the problem. Multi-agent systems — multiple LLMs interacting with one another, often proposed as an antidote to the limitations of single models — do not necessarily eliminate sycophantic drift but may transfer it to the level of inter-agent interaction. An analysis of failure modes in multi-agent debate

(Wynn et al., 2025) documented how debates between LLMs can collapse into false disagreement or premature convergence rather than producing genuine challenge; a controlled study on logical reasoning tasks (Wu et al., 2025) found that debate produces limited benefits dependent on task structure. The multiplication of agents is therefore not in itself a solution to structural sycophancy: the adversarial mandate must be incorporated into the system architecture, not merely into the interface.

2.2 — AUTOMATION BIAS AND OVER-RELIANCE: THE SURRENDER OF CRITICAL CONTROL

Sycophancy would be less serious if decision-makers systematically resisted it, integrating AI output as one piece of information among many, critically evaluated against their own expert judgment. The evidence on human–automation interaction suggests that this assumption is optimistic.

Parasuraman and Manzey (2010) developed an integrated model of complacency in automation use, synthesizing decades of research on semi-automatic systems in aviation, medicine, industrial control, and complex process supervision. Their framework distinguishes *automation complacency* — the reduction of vigilance when the system is perceived as reliable — from *automation bias* (here and subsequently also *over-reliance*): the systematic tendency to treat the automated recommendation as the focal point of reasoning, attributing to it a weight exceeding that justified by the system's actual quality. Both phenomena manifest independently of the system's real accuracy: the perception of reliability, once established, resists updating even in the face of detectable anomalies.

The implications for leadership decision-making processes are direct. When an AI system produces an articulate, structured analysis that appears to be grounded in evidence, the decision-maker is subject to the same dynamics that Parasuraman and Manzey describe for operators of automatic systems: treating the output as the privileged starting point, reducing the search for contrary information, and lowering critical vigilance to the extent that the system appears reliable. In high-stakes contexts — where automation errors are rare but consequential — this reduction of attention is most dangerous: it generates failures concentrated in the cases where the system errs with the greatest apparent confidence.

2.3 — DESKILLING: THE SILENT DETERIORATION OF EXPERTISE

A third mechanism operates over longer time horizons and is therefore harder to observe in real time: the prolonged use of automated systems erodes the independent judgment skills that those systems are meant to support. The classic formulation remains that of Bainbridge (1983) in the essay «Ironies of Automation», still relevant more than four decades on.

Bainbridge identified a structural paradox in the engineering of automated systems: designers assign the operator the task of intervening when automation fails — in the most critical and atypical cases — but the automation itself, by reducing the frequency with which the necessary skills are exercised, atrophies the capabilities required for that intervention. The more reliable the system in normal use, the less the operator practices the skills required under fault conditions, and the less capable they will be of intervening when needed. The irony is twofold: the advance of automation does not simplify the cognitive role of the operator but transforms it, requiring a peak of competence precisely when the practice of that competence has been minimized by the automation itself.

The transposition to leadership contexts is not metaphorical. An executive who systematically delegates to AI assistants the synthesis of competitive contexts, risk assessment, and the construction of alternative scenarios does not stop deciding, but reduces the frequency with which they independently build those models of the world. In the short term the impact is invisible; in the medium term, the capacity to recognize the limits of AI analysis, to interrogate its premises with disciplinary competence, and to produce independent judgments when the tool fails — typically in cases outside the distribution on which it was trained — deteriorates silently. Deskilling does not cancel judgment: it progressively hollows out its critical content.

2.4 — COGNITIVE HOMOGENIZATION: THE RISK OF ALGORITHMIC MONOCULTURE

The three mechanisms described operate primarily at the individual or organizational level. A fourth, of a systemic nature, emerges when considering the aggregate of decision-makers in a sector or institution: the convergence of AI tools can produce a convergence of reasoning whose collective consequences are not reducible to the sum of individual effects.

Kleinberg and Raghavan (2021) formally analyzed the implications of *algorithmic monoculture* — the condition in which many organizations adopt the same decision-support algorithm — in models of selection and allocation. Their analysis shows that the systemic outcomes of monoculture can differ from those of diversified adoption, even when each individual adopter benefits individually. The mechanism is the correlation

of errors: when many agents use the same predictive model, their systematic errors cease to be independent and become covariant, eroding the diversification benefit that in conditions of heterogeneity absorbs the erroneous evaluations of individuals.

Transposed to AI applied to leadership decisions, the result has immediate relevance. When investment committees, boards of directors, and analysts of different organizations turn to the same AI systems, they are not only individually ceding judgment autonomy: they are contributing to a structure of correlations that reduces the systemic diversity of available reasoning. Divergent evaluations — which in healthy markets and deliberative processes surface information that the dominant reading overlooks — become rarer, not through explicit agreement but through convergence of tools. Cognitive homogenization is not a merely future threat: it is already, in sectors with high AI penetration, a structural condition in progress.

2.5 — THE GENEALOGY OF STRUCTURED CHALLENGE: FROM DEVIL'S ADVOCACY TO ADVERSARIAL AI

The four mechanisms are not without conceptual antidotes. The literature on structured challenge offers a well-established genealogy, which predates modern language models entirely and must be examined to situate the contribution of this paper within the continuity of ideas.

Schwenk (1990) examined the experimental and field evidence on *devil's advocacy* and *dialectical inquiry* as prescriptive techniques for improving strategic decisions. His review documents that groups in which an agent has the institutional mandate to argue against the prevailing proposal produce, on average, evaluations of superior quality compared to those in which dissent is merely permitted but not structurally required. The central finding is not that disagreement is useful as such, but that its institutionalization — its incorporation into the architecture of the process rather than into the character of the participants — appears to be a necessary condition for its effectiveness.

That tradition found, four decades later, an echo in the work of Cai, Arawjo, and Glassman (2024), who formalized the design space for AI systems whose utility function is explicitly oppositional — the so-called *antagonistic AI* — articulating the tensions between systems oriented toward user satisfaction and systems oriented toward a more demanding epistemic utility that includes the active challenge of the decision-maker's reasoning. The articulation of this space is conceptually necessary before one can discuss an AI system that operationalizes devil's advocacy in a reliable and not merely performative manner.

The gap that this genealogy leaves open is precise. Neither the literature on structured challenge nor that on antagonistic AI offers a theoretical construct that links the four erosion mechanisms — sycophancy, over-reliance, deskilling, cognitive monoculture — to a unitary notion of *independent judgment capacity* that is recognizable, monit-

orable, and protectable at the governance level. Devil's advocacy is a procedural technique, antagonistic AI a design space, algorithmic monoculture a systemic risk: each elaboration remains circumscribed to its own domain, without an integrated framework that describes the cognitive condition of leadership in the age of AI and specifies the conditions for its protection. It is this unifying construct — *cognitive sovereignty* — that §3 introduces and develops.

SECTION 3

The construct: cognitive sovereignty

The gap identified in §2.5 is conceptual: the four erosion mechanisms — sycophancy, over-reliance, deskilling, monoculture — have been studied as separate phenomena, each in its own literature, without a construct that links them to a single threatened capacity. This section proposes it as a theoretical proposal to be operationalized, not as a quantity already measured or empirically validated: its definition, the articulation of dimensions, the analysis of erosion, and the conditions of observability follow.

3.1 — FORMAL DEFINITION

We define the **cognitive sovereignty** of a leadership body as *the collective capacity to maintain independent, calibrated, and falsifiable judgment on consequential decisions, in the face of artificial intelligence systems that structurally tend to please the user and to homogenize reasoning among decision-makers*. The definition contains three non-negotiable qualifiers. Judgment is *independent* when its direction is not determined by the conformative pull of AI output or by the consensus of tools adopted by peers. It is *calibrated* when the confidence estimates that accompany it track the empirical frequencies of outcomes, in the sense of the literature on accurate forecasting (Tetlock & Gardner, 2015). It is *falsifiable* when it is formulated in a way that specifies what evidence would refute it and on whom the burden of producing it falls.

This notion must be distinguished from two adjacent concepts. It does not coincide with *skepticism*: skepticism is the disposition to suspend assent and, elevated to a rule, produces paralysis or indiscriminate rejection of evidence, while cognitive sovereignty requires calibrated assent — believing to the extent justified by the evidence, neither more nor less. Nor does it coincide with generic *autonomy*: one can decide without external constraints while maintaining a poorly calibrated or conformist judgment. Cognitive sovereignty is more stringent — it is autonomy *of a judgment that remains independent, calibrated, and falsifiable under the specific pressure exerted by AI systems*: a property of judgment under adversarial load, not of the decision-maker in the abstract.

Two clarifications complete the definition. Cognitive sovereignty is an attribute of a *leadership body* — a board, a committee, a deliberative function — not of the individual: it is a property of the collective process through which an organization forms and tests its judgments, and is eroded or protected at that level. And it is *gradual*, not binary: it is not possessed or lost, but maintained to a greater or lesser degree along the dimensions that follow.

3.2 — DIMENSIONS

We propose four dimensions of the construct. Each has an operational definition — the quantity that, in principle, would make it observable — and responds to a specific erosion mechanism from §2. They are proposed as an articulation of the construct to be operationalized and validated, not as scales already constructed and measured.

(1) Independence of judgment. This is the resistance of judgment to the conformative pull of AI output: the extent to which the decision-maker's final position diverges, when the evidence justifies it, from that which the AI assistant has validated or suggested. Operationally it manifests in the discrepancy between the judgment formed before and after exposure to AI output, conditioned on the quality of the evidence: a decision-maker with high independence updates in the direction of the evidence, not of the output. It responds directly to the sycophancy documented in RLHF models (Sharma et al., 2023): sycophancy is the vector that pushes judgment toward the pleasing position, and independence is its countermeasure at the capacity level.

(2) Resistance to homogenization. This is the divergence preserved from model consensus: the extent to which an organization's judgment remains distinct from what others, fed the same AI systems, would form from the same evidence. Operationally it is observed in the correlation between the judgments of different decision-makers who share the same tool — low correlation (holding evidence constant) indicates preserved resistance, high correlation indicates homogenization. It responds to the risk of algorithmic monoculture (Kleinberg & Raghavan, 2021): where convergence of tools makes systematic errors covariant, this is the capacity that maintains the diversity of judgment on which healthy deliberative processes depend.

(3) Uncertainty calibration. This is the property by which confidence estimates track the empirical frequencies of outcomes: among evaluations expressed with 70% confidence, approximately 70% should prove correct ex post. Operationally it is the quantity measured by calibration curves and scoring rules from the forecasting literature (Tetlock & Gardner, 2015). It responds to the combined effect of two mechanisms from §2: sycophancy inflates confidence by providing undeserved validation (Sharma et al., 2023), while over-reliance induces attributing to automatic output a trust exceeding its actual reliability (Parasuraman & Manzey, 2010). Calibration is the dimension that both these vectors erode from the side of uncertainty estimation.

(4) Burden-of-proof traceability. This is the property by which, in a decision, it is explicit and recorded who must demonstrate what: which party holds the thesis, which has the mandate to refute it, which assertions remain unproven, and what evidence would confirm or falsify them. Operationally it is observed in the presence, within the deliberative process, of a verifiable record that attributes the burden of proof and registers its fulfillment or non-fulfillment. It responds to over-reliance (Parasuraman & Manzey, 2010) and deskilling (Bainbridge, 1983): when AI output becomes the unquest-

ioned focal point of reasoning, the burden of proof dissolves silently — no one is required to demonstrate anything, because the tool «has already analyzed» — and the very capacity to formulate what proof would be needed atrophies with disuse.

3.3 — ANTECEDENTS AND EROSION

The four dimensions do not weaken independently: the mechanisms of §2 are causal antecedents that erode them through distinct and often simultaneous pathways. Instrumental sycophancy (§2.1) attacks independence (1) and calibration (3); over-reliance (§2.2), by dissolving the attribution of burden of proof and inducing poorly calibrated trust, attacks traceability (4) and calibration (3); deskilling (§2.3), by eroding the practice of autonomous judgment, attacks independence (1) and traceability (4); cognitive homogenization (§2.4) operates systemically on resistance to homogenization (2), making the judgments of organizations that share tools covariant.

Two properties of this erosion deserve emphasis, because they explain why cognitive sovereignty requires deliberate protection rather than being assumed spontaneously stable. The first is *silence*: none of the four mechanisms produces a warning signal while acting. Sycophancy presents itself as articulate agreement, over-reliance as efficiency, deskilling is invisible in the short term (§2.3), and monoculture emerges only at the aggregation of many decisions. The tendency to seek confirmation of preferred hypotheses (Kahneman, 2011) means that the decision-maker does not perceive erosion as loss but as confirmation of their own judgment. The second is *covariance*: the mechanisms do not add but reinforce one another. Sycophancy feeds over-reliance — an output that validates is an output to be trusted — which accelerates deskilling, which in turn reduces the capacity to resist sycophancy. It is this dynamic of mutual reinforcement, not a single isolated mechanism, that makes the cognitive condition of leadership structurally exposed.

3.4 — OBSERVABILITY AND MEASUREMENT

A construct is scientifically useful to the extent it specifies the conditions under which it becomes observable. Cognitive sovereignty, as defined in §3.1, is not directly observable: it is a capacity, not an event. It becomes observable only through its manifestations along the four dimensions, and under three conditions. The first is a *decision trail*: dimensions (1), (3), and (4) require that judgment prior to exposure to AI output, declared confidence, and attribution of burden of proof be recorded at the moment of decision, not reconstructed ex post — a prospective deliberative register is the precondition for any observation. The second is *outcome resolution*: calibration (3) cannot be assessed until a sufficient number of judgments has been compared against outcomes. The third is *comparison across decision-makers*: resistance to homogenization (2) is a relat-

ive quantity, observable only by comparing the judgments of multiple organizations holding evidence constant, not measurable on a single isolated decision-maker.

Within this framework, the **Δ-CSI** (Cognitive Sovereignty Index, delta) introduced in *Calibrated Dissent* (Canepa, 2026) is *one* of the possible tools for anchoring the construct — not the only one, and not a measure of cognitive sovereignty as a whole. By the operational definition given in that work (Canepa, 2026, §2.3), it is a **proxy of the challenge intensity applied** in a session — how forcefully the pressure was exerted on the reasoning — **and not a measure of decision correctness**; treating it as an accuracy signal is a category error. Among the four dimensions it primarily anchors burden-of-proof traceability (4), of which it records session by session the activity of refutation; it does not measure independence (1), resistance to homogenization (2), or calibration (3), which require respectively pre/post-exposure comparison, comparison across decision-makers, and outcome resolution. Cognitive sovereignty is therefore not reducible to the Δ-CSI: constructing instruments for the other three dimensions — and validating them empirically — remains an open research program, not a result of this paper, which proposes the construct and specifies its conditions of observability without claiming to have measured it.

SECTION 4

The mechanism: design principles of the cognitive contradictor

The preceding section identified the four dimensions of the construct and their respective erosion antecedents. Recognizing the construct is not enough: a mechanism is needed that operates actively to protect it. It is the **cognitive contradictor** — a system designed not to assist the decision-maker's reasoning, but to challenge it. The challenge is the utility function of the system, not an occasional by-product of an analysis oriented toward synthesis.

The term *cognitive contradictor* does not designate a specific product nor a patented architecture: it designates a class of systems that share a set of necessary design principles — four, as we argue below — without which a system cannot claim to be a contradictor in a substantive sense, whatever label it adopts. They are not optional configurations, but architectural invariants that distinguish the contradictor from the sophisticated validator. As the literature on devil's advocacy documents, the effectiveness of structured challenge depends on its institutionalization — on its incorporation into the architecture of the process, not into the character of the participants (Schwenk, 1990); the same applies to an AI system that aims to operationalize that mandate at scale.

The four principles are reusable by anyone designing AI systems with an adversarial mandate. Each responds to one or more erosion mechanisms from §2 and protects specific dimensions of the construct from §3. A reference implementation satisfying all four — CounterBrain — is discussed at the end of the section with the sole purpose of showing that the principles are not purely theoretical but achievable in an operational context.

4.1 — STRUCTURAL ADVERSARIALITY

The first principle establishes that the system's mandate is to challenge the decision, never to validate it. This is more stringent than it appears: it does not require that the system produce *also* a challenge alongside the analysis, but that the challenge *is* the analysis. Validation is not a secondary objective to be balanced against opposition: it is a mode incompatible with the contradictor's role.

Cai, Arawjo, and Glassman (2024) formalized the design space for antagonistic AI, articulating the gradient between systems oriented toward user satisfaction and systems whose objective is explicitly oppositional. Structural adversariality places the contradictor at the oppositional extreme of that gradient without intermediate configurations: there is no «moderate contradictor» or «softened challenge» mode. This is not an

ensity is configurable would return, by organizational inertia or at the client's will, toward the validating extreme — precisely the optimization that training for user satisfaction structurally incentivizes (Sharma et al., 2023).

The experimental evidence reinforces this position. Schwenk (1990) documented that groups in which devil's advocacy is *required* produce strategic decisions of superior quality compared to those in which it is merely *permitted*, because permissiveness allows the challenge to be omitted precisely when it is most necessary — when the group is already convinced of its thesis. The structural mandate removes that escape route. For an AI system the corollary is direct: if the adversarial mode is configurable, it will be disabled precisely by the decision-makers most exposed to sycophancy.

The dimension this principle primarily protects is **independence of judgment** (§3.2, dimension 1): a system whose mandate is to validate cannot, by definition, preserve the non-conformative character of judgment. Sycophancy compresses the distance between the decision-maker's judgment and the position validated by the tool (Sharma et al., 2023); structural adversariality opposes an opposite, systematic, and non-modulable force. The same mandate also protects **resistance to homogenization** (dimension 2): by forcing divergence from the dominant reading rather than convergence toward model consensus, it preserves the distance of judgment between the organization and peers sharing the same tools.

4.2 — FAIL-CLOSED ON SYCOPHANCY

The second principle specifies the failure mode of the system. If the output of a pass does not meet the mandatory challenge threshold — if the system is about to emit an analysis that does not genuinely challenge the thesis — it must raise an explicit error, not silently emit the non-challenging output. Sycophancy is a failure mode, not a degraded one: the functional equivalent of a type-II error, where the non-rejection of a false hypothesis is a test failure, not a neutral result.

An empty or minimal challenge that passes as valid — in the absence of an explicit gate — is indistinguishable from the output of a robust contradictor by the observer who cannot inspect the pipeline. The decision-maker is then falsely reassured by the *form* of the output (it has the appearance of a structured challenge) while the content lacks adversarial force: a more serious failure mode than the absence of challenge, because it removes the perception of risk without eliminating its substance.

Research on multi-agent systems documents corresponding phenomena: debates between LLMs that collapse into false disagreement or premature convergence (Wynn et al., 2025) and reasoning tasks in which debate produces limited benefits dependent on task structure (Wu et al., 2025). The multiplication of agents is therefore not in itself a protection against fail-open: an explicit gate is needed that verifies, after each adversarial pass, that the challenge is non-empty across each of the mandatory dimensions,

with a dedicated retry and, on second failure, an engine error that halts the pipeline without emitting partial outputs.

The fail-closed principle primarily protects **burden-of-proof traceability** (§3.2, dimension 4). The corresponding mechanism is over-reliance (Parasuraman & Manzey, 2010): when an AI output appears structurally complete, the decision-maker tends to treat it as a focal point, reducing attention to verifying adversarial strength. Fail-closed ensures that this trust is not accorded to outputs that do not merit it; the protection is architectural, not dependent on the decision-maker's vigilance.

4.3 — FIXED OUTPUT SCHEMA

The third principle establishes that the contradictor's output conforms to a schema with mandatory fields that no vertical specialization or client configuration can remove, rename, or make optional. The fixed structure is not a stylistic preference: it is the architectural translation of the invariance of challenge, its externalization into a verifiable contract.

Calibrated Dissent (Canepa, 2026, §2.2) specifies a seven-field schema — construction, properties, and motivations documented field by field there — versioned as a JSON contract with `additionalProperties: false`. We do not reproduce it here: what matters is the *principled function* of the fixed schema, independent of the specific values adopted for the seven fields.

That function is threefold. First, each mandatory field corresponds to a specific manifestation of construct erosion: implicit assumptions make visible the premises that sycophancy would leave silently confirmed; falsification tests translate the burden of proof into actionable empirical checks; calibrated confidence, with declared knowledge gaps, counters the inflation of trust induced by over-reliance (Parasuraman & Manzey, 2010). Second, the fixity of the schema enables longitudinal comparison and quality control: if a field can be omitted, the guarantee that the challenge of one session is comparable to previous ones is lost, and the accumulation of data for calibration remains without foundation. Third, the non-renamable character of the fields prevents lexical drift — the risk that an alternative configuration reintroduces the same fields with a vocabulary that weakens the adversarial mandate (for example renaming «falsification tests» as «complementary observations»).

The fixed schema protects all four dimensions transversally, with particular emphasis on **uncertainty calibration** (dimension 3) — the mandatory presence of calibrated confidence and explicit knowledge gaps is the necessary condition for estimates to be comparable against realized outcomes — and on **burden-of-proof traceability** (dimension 4), for the structure of the falsification fields and burden-of-proof questions as distinct and mandatory elements.

4.4 — PROVENANCE/SCORING SEPARATION

The fourth principle concerns a circular dependency that retrieval-enabled analysis systems can introduce in the absence of explicit design: the material retrieved as provenance must not retroactively alter the quantities subjected to judgment.

The risk is as follows. A contradictor equipped with search and retrieval tools may find, during analysis, evidence that modifies the formulation of the challenge. If that same evidence simultaneously influences the content of the output and the measurement of intensity — specifically the Δ-CSI, proxy of the challenge intensity applied (Canepa, 2026, §2.3) — a circularity is introduced in which provenance and scoring co-determine each other: high Δ-CSI values may then be an artifact of the availability of abundant sources, not of an adversarial intensity genuinely applied to the reasoning. The index ceases to measure the property it is intended to monitor.

The separation is realized architecturally by isolating the retrieval phase from the scoring phase: sources are retrieved and recorded as provenance, but the Δ-CSI components are calculated deterministically from the logical structure of the challenge — assumptions, falsification tests, burden-of-proof questions, scenario divergence — not from the volume or quality of the retrieved material. The count of sources contributes to the index as a separate component with its own weight, but does not retroact on the others nor alter the logical structure of the challenge already produced. This is the condition for the Δ-CSI to remain interpretable as a proxy of challenge intensity, invariant with respect to the information context, and not as a measure of decision correctness — a distinction the paper treats as an interpretive invariant (§3.4; Canepa, 2026, §2.3).

The separation principle primarily protects **uncertainty calibration** (dimension 3): the empirical validity of a challenge intensity index requires that it measure a property of reasoning, not of retrieval. A Δ-CSI conflated with the abundance of sources cannot be calibrated in the sense of the forecasting literature — as correspondence between declared confidence estimates and empirical frequencies of outcomes (Tetlock & Gardner, 2015).

COUNTERBRAIN — REFERENCE IMPLEMENTATION

The four principles of this section — structural adversariality, fail-closed on sycophancy, fixed output schema, provenance/scoring separation — are necessary conditions for any system intending to operate as a cognitive contradictor; they are neither proprietary nor bound to a specific architecture or organization. CounterBrain, the system whose measurement protocol *Calibrated Dissent* (Canepa, 2026) reports, is the reference implementation that demonstrates the achievability of these principles in an operational context: adversariality coded as a non-configurable invariant; fail-closed as an explicit gate with retry and pipeline interruption in the absence of a valid challenge; a seven-field schema versioned as a JSON contract with `additionalProperties: false`; provenance/scoring separation architecturally guaranteed (Canepa, 2026, §2.4). Alternative systems satisfying the same four principles would equally fulfill the contradictor mandate. The mention serves a single epistemic purpose: to situate the principles in the domain of the operationalizable, not the theoretical ideal.

SECTION 5

Illustrative case: the Osservatorio (N = 10)

> METHODOLOGICAL CAVEAT
>
> **ILLUSTRATES, DOES NOT PROVE.** The material presented in this section illustrates that the device outlined in §3 and §4 has been effectively instantiated on real decisions. It does not in any way constitute a test of the effectiveness of the cognitive contradictor, does not permit evaluation of whether the forecasts are calibrated, and does not offer confirmatory evidence for any hypothesis of the research program. The confirmatory test is reserved for a powered Phase 2, pre-registered in the companion paper (Canepa, 2026); this section is the account of the start of the device, not of its verdict.

5.1 — THE DEVICE: PUBLIC REGISTER, APPEND-ONLY, HASH-ANCHORED

The construct of cognitive sovereignty (§3) requires a prospective deliberative register: judgment must be fixed before any outcome is observable, in a form that precludes retrospective revision — without which calibration measures only an illusion constructed ex post. The contradictor mechanism (§4) equally requires that challenge be structural, that output respect a fixed and non-renamable schema, and that provenance not retroactively alter the quantities submitted to scoring (§4.4). The operational question is whether a device satisfying these conditions simultaneously is achievable outside the laboratory, on real and consequential decisions, with non-manipulability assurances independently verifiable by third parties.

The **Osservatorio** is the concrete answer to that question: a public, append-only register in which each forecast card — produced by the contradictor process — is published before any outcome is known. Integrity is guaranteed by a double anchor. The first is a SHA-256 hash of the content at the moment of sealing: any subsequent modification produces a different hash, making revision immediately detectable. The second is an OpenTimestamps (OTS) anchor, which embeds the hash in a transaction on the Bitcoin blockchain, whose immutability depends on the distributed computational power of the entire network, not on trust in a single operator. The SHA-256 seal + OTS anchor pair fulfills the burden-of-proof traceability requirement (§3.2, dimension 4): any external reviewer can independently verify that the published text corresponds to the original hash and that that hash was recorded on the blockchain before a given instant.

The sealing and resolution protocol — the deterministic rules for establishing when an outcome has materialized, the canonical sources, the recommended double blind coding — is documented in Canepa (2026, §4–§5). This section adds only the observation that the protocol has effectively been started on real decisions.

5.2 — THE PILOT: N = 10 FORECASTS SEALED ON 20 JUNE 2026

On **20 June 2026** (To), **ten forecasts on ten high-stakes European merger and acquisition operations** were sealed in the Osservatorio. Each public card records: the three-way scenario distribution produced by Pass 1 of the contradictor (dominant reading), the stated main thesis, a P(H1) estimate — deal-break or downward renegotiation ≥ 15% over the 6–12-month horizon — and, for the pre-declared subset of four operations with a listed acquirer consolidating the target, a P(H2) estimate for the 18–36-month horizon. The complete technical payload of the engine — seven-field schema, Δ-CSI, and bundle hash freezing model, prompt, and Pack — is preserved as provenance not submitted to scoring (§4.4) and does not appear in the public cards at the moment of sealing.

The pre-registration of the cohort is public: **osf.io/5qc8t**, DOI 10.17605/OSF.IO/5QC8T (Canepa, 2026). The OSF deposit was made on 25 June 2026 without embargo: the pre-specified material is freely accessible at the publication of this paper. The Osservatorio is accessible at counterbrain.com/registro, where sealed cards and the progressive resolution trail are available without authentication.

Resolution windows extend from June 2027 (H1, 12 months) to June 2029 (H2, 36 months). At the time of writing this paper no outcome is yet observable for the full sample.

5.3 — EPISTEMIC PERIMETER: WHAT THE CASE ILLUSTRATES AND WHAT IT DOES NOT CLAIM

ILLUSTRATES, DOES NOT PROVE. The case answers a single question: are the design principles of the cognitive contradictor (§4) achievable in an operational context on real decisions? The answer the case permits is affirmative in the sense of existence and operability: structural adversariality was applied to consequential forecasts, not to hypothetical laboratory stimuli; the seven-field schema was produced and versioned; the fail-closed pipeline operated on a defined set of cases; provenance/scoring separation is implemented as an architectural invariant; the append-only register with OTS anchoring exists and is verifiable by anyone. On the construct side (§3), the condition of prospective observability — the recording of judgment before the outcome — is satisfied, and burden-of-proof traceability (dimension 4) is recorded for each card.

This is the exact epistemic perimeter. Beyond that boundary, the case claims nothing. It does not claim that the contradictor is effective. It does not claim that the forecasts are calibrated. It does not claim that the design principles produce decisions of superior quality. It does not claim that the Δ-CSI predicts the correctness of the challenged decisions: the Δ-CSI, where cited in the provenance cards, is exclusively a proxy of the challenge intensity applied in that session (§4.4; Canepa, 2026, §2.3), meaning it

measures how forcefully the process pressed on the dimensions of the construct, not whether the challenged decision has good chances of success. Treating it as an accuracy signal is a category error (§3.4). N = 10 is structurally below the power threshold required for any confirmatory test — fixed at N ≥ 20 per horizon in Canepa (2026, §6.5) — and outcomes are largely still in the future.

The confirmatory test — the evaluation of calibration (H-cal) and divergence-that-pays (H-pay) against the pre-registered baselines — is reserved for a **powered Phase 2**, with a larger, forward-sealed cohort, whose analysis conditions are pre-registered in Canepa (2026, §6.5). That phase will produce confirmatory evidence or, if the hypotheses do not hold, their falsification. The Osservatorio is the device that makes an honest test without hindsight possible; it does not anticipate its verdict.

SECTION 6

Governance: protecting cognitive sovereignty at board level

The problem (§2) defined the mechanisms through which AI degrades leadership judgment; the mechanism (§4) specified the principles a cognitive contradictor must respect to operate in a genuinely adversarial manner. The third phase of the arc — adoption — addresses a distinct question: how can a leadership body institutionalize the challenge function in its own deliberative architecture? The answer is not technical but organizational: it requires explicit protocols, role attributions, decision registers, and calibration mechanisms that operate structurally, not through episodic individual vigilance.

6.1 — FRAMING: REGULATING ITS USE, NOT PROHIBITING IT

The temptation to address the risks of §2 by limiting AI in high-stakes decisions is understandable but misdirected. Model sycophancy (Sharma et al., 2023), over-reliance (Parasuraman & Manzey, 2010), and deskilling (Bainbridge, 1983) do not depend on the intensity of use but on its architecture: the same system used without structured challenge produces erosion; with an institutionalized contradictor it produces epistemic pressure. The problem is not AI in consequential decisions — it is AI without a contradictor.

The lesson from the literature on structured challenge is that the determining variable in the quality of strategic decisions is not the competence of the participants nor the quantity of information, but the architecture of the process (Schwenk, 1990). The governance of the contradictor must inherit this: the challenge mandate must be codified in that architecture, not delegated to the character or vigilance of individuals.

6.2 — FOUR GOVERNANCE PRINCIPLES

The following four principles translate the design logic of §4 to the organizational level. They are formulated as practices of institutional prudence for deliberative bodies operating under conditions of high AI assistance — not as a derivation from specifically codified regulatory obligations, but as a rational response to the documented erosion dynamics.

(i) Mandatory challenge at the decision point. Every decision that exceeds a pre-defined materiality threshold must activate a challenge mandate before the body acts. Mandatoriness is the core of the principle: a challenge that can be omitted tends to be in cases of greatest pressure — precisely those in which it would be most necessary (Schwenk, 1990). The threshold must be defined ex ante by a function with no interest

in the decision, on criteria of stake value, reversibility, and strategic scope; reversibility is the most critical, because a decision that is not reversible at reasonable cost within a predefined horizon is a candidate for challenge regardless of monetary value.

(ii) Recording and auditing of the decision and the challenge. A consequential decision must generate a deliberative trail recording: the question formulated before recourse to AI; the output of the challenge process, with critical assumptions, proposed falsification tests, and confidence with declared knowledge gaps; the body's deliberation and the factors that determined it; the decision taken. The burden of identifying critical assumptions and falsification tests lies with the challenge function; the body explicitly states why the decision holds despite the challenge. The trail is produced at the moment of decision, not ex post: without prospective recording, post-outcome rationalization — reinforced by the same sycophantic tendencies of AI models (Sharma et al., 2023) — renders empirical calibration unreachable (Tetlock & Gardner, 2015).

(iii) Role separation: who decides ≠ who challenges. The challenge function must not be entrusted to the same person or unit that prepared the original analysis. This is the organizational parallel of structural adversariality (§4.1): if the challenger has an interest in the adequacy of the analysis they produced, the challenge tends toward validation for the same motivational reasons that induce RLHF models toward sycophancy (Sharma et al., 2023). Separation is achieved through rotation of responsibilities, external appointment for specific categories of decisions, or — in AI-assisted implementations — architectural isolation of the challenge function; what it excludes is the arrangement in which whoever produces the analysis also judges whether it has been adequately challenged.

(iv) Verified calibration over time. A body that applies structured challenge systematically accumulates a register of decisions, confidence estimates, and, progressively, outcomes. At regular intervals — at minimum annually, or upon reaching a sufficient number of resolved decisions — that register must be examined: were the estimates calibrated in the sense of Tetlock and Gardner (2015)? Were the falsification tests actionable or merely formal? Do systematic failures emerge in specific categories? This is not an evaluation of individual performance, but a calibration audit of the process, designed to detect the slow drift — the deskilling described by Bainbridge (1983) — before it renders the quality of decisions unassessable.

6.3 — THE DELIBERATIVE REGISTER AS AN ORGANIZATIONAL PATTERN

Principles (ii) and (iv) converge toward a common architecture: the **deliberative register**. The requirement is simple but stringent — the decision trail must precede the outcome, be not retroactively modifiable, and be accessible to the calibration function. This is a functional specification, not a technology: any implementation that ensures

prospective recording, append-only structure, and independent verifiability satisfies the requirement.

This pattern is sometimes called the *Sovereign Decision Ledger* — a term designating the organizational pattern, not a specific system. The principle is that judgment, challenge, and attribution of burden of proof be fixed before any outcome is observable, in a form that precludes silent revision, without which burden-of-proof traceability (§3.2) remains nominal. The Osservatorio of §5 is an operational instantiation of it, with architecture documented in Canepa (2026). The pattern is technology-independent and scalable: applying it requires procedural discipline, not specialized infrastructure.

The minimum condition for the register to have epistemic value is that the trail be produced before the decision-maker knows the outcome, even partial: any feedback of the outcome on the attribution of burden of proof invalidates calibration and transforms the register into documented rationalization.

6.4 — CONNECTION TO FIDUCIARY DUTIES: A MATTER OF EPISTEMIC PRUDENCE

The governance principles of this section do not derive from a specifically codified regulatory framework — it is not claimed that a statutory prescription requires them. They are, however, consistent with the logic of fiduciary duties that govern deliberative bodies in most legal systems. The duty of care requires that decision-makers act on an informed basis, in good faith, with the care of an ordinarily prudent person. Operating under documented conditions of over-reliance (Parasuraman & Manzey, 2010) — in which AI output systematically becomes the focal point of reasoning, with a weight exceeding the system's actual reliability — without a structural challenge mechanism raises questions about compatibility with that standard of prudence, on the basis of its own cognitive presuppositions and not by effect of an external prescription.

From this perspective, mandatory challenge, the deliberative trail, and periodic calibration are not procedural ornaments: they are the structural conditions that make the standard of informed deliberation required by institutional prudence accessible. Deskilling (Bainbridge, 1983) and cognitive homogenization (Kleinberg & Raghavan, 2021) operate over horizons that individual bodies do not perceive in real time: addressing them requires institutional mechanisms, not episodic individual vigilance.

The effectiveness of these principles applied specifically to AI-assisted deliberation at board level has not been tested in a controlled manner; the available evidence comes from the organizational literature (Schwenk, 1990) and from calibration research (Tetlock & Gardner, 2015), whose transferability to this specific context requires caution. The principles offer a coherent translation of the logic of §4 to the institutional level; the deliberative register produced by principle (iv) is the mechanism through which an organization can accumulate, over time, evidence on the effectiveness of its challenge device.

SECTION 7

Limitations and what this paper does not claim

This paper proposes a theoretical construct and the design principles of the corresponding mechanism; it states its epistemic boundaries explicitly here, distinguishing what it claims from what could be mistakenly read as implied.

On the effectiveness of the cognitive contradictor. This paper does not claim that the cognitive contradictor improves the quality of decisions. The four design principles of §4 — structural adversariality, fail-closed on sycophancy, fixed output schema, provenance/scoring separation — are necessary conditions for a system aspiring to operationalize devil's advocacy in a non-merely-performative manner; they are not sufficient conditions to guarantee decisions of superior quality, nor a demonstrated mandate of causal effectiveness. Effectiveness is an empirical question this paper does not resolve and does not pretend to resolve.

On cognitive sovereignty as a predictor of correctness. This paper does not claim that cognitive sovereignty predicts the correctness of decisions. The four dimensions of the construct (§3.2) describe the quality conditions of the deliberative process, not the soundness of its outcomes. A leadership body can preserve its cognitive sovereignty intact and still make erroneous decisions; it can erode it systematically and, due to favorable circumstances, obtain positive results. Cognitive sovereignty is a property of the process, not a predictor of the outcome. Similarly, the Δ-CSI — proxy of the challenge intensity applied in a session (Canepa, 2026, §2.3) — does not predict correctness: a high value signals that the process pressed hard on many dimensions, not that the decision has good chances of success. Treating it as an accuracy signal is a category error (§3.4).

On the empirical status of the construct. The construct of cognitive sovereignty is a theoretical proposal, not a measured or empirically validated quantity. The four dimensions are articulated with operational definitions that specify the conditions of their observability; they are not scales already constructed, tested, and psychometrically validated. The operationalization program — the translation of operational definitions into testable measurement instruments, the test of factorial structure, the estimation of convergent and discriminant validity with respect to adjacent constructs — is a future research agenda, explicitly declared in §8. The Δ-CSI anchors only one dimension (burden-of-proof traceability, §3.4) and is by definition a proxy of challenge intensity, not a measure of the construct as a whole.

On the illustrative case N = 10. The Osservatorio pilot presented in §5 is non-evidence for any claim of contradictor effectiveness. As stated explicitly in §5.3, the case makes available evidence of operational instantiation: the design principles of §4 were

applied to real decisions and the condition of prospective traceability is satisfied; no evidence is derived that the contradictor produces calibrated forecasts, that it distinguishes itself from a base-rate baseline, or that it affects the quality of challenged decisions. N = 10 is structurally below the power threshold required for any confirmatory test, fixed at N ≥ 20 per horizon in Canepa (2026, §6.5). Outcomes are largely still in the future; the confirmatory test is reserved for the powered Phase 2 pre-registered in that work.

On generalizability. The construct is developed with primary reference to high-stakes, low-reversibility decision contexts — M&A, institutional investment, board strategic deliberations. It is not claimed to be transferable without re-examination to low-stakes domains, high-frequency decision-making, or contexts with a radically different incentive structure. Whether the erosion mechanisms (§2) manifest in the same forms and whether the design and governance principles counter them with the same effectiveness in other domains is an open empirical question.

On the interpretation of the literature. The erosion mechanisms of §2 and the principles of §4 rest on literature with independent scientific standing — Schwenk (1990), Bainbridge (1983), Sharma et al. (2023), Kleinberg and Raghavan (2021). The interpretation provided, and especially its application to the condition of AI-assisted leadership, is a theoretical contribution awaiting validation; the paper does not claim it is the only possible reading nor that the cited literature exhausts the relevant landscape.

Conclusion and research agenda

This paper has argued along an arc in three moves: problem, mechanism, adoption.

The problem (§2) is structural: AI systems optimized for user satisfaction erode leadership judgment through four distinct and mutually reinforcing mechanisms — instrumental sycophancy, over-reliance, progressive deskilling, and cognitive monoculture — that operate silently and worsen under high AI assistance. The open gap is the absence of a unifying construct linking them to a collective capacity — the sovereignty of judgment — that is recognizable, monitorable, and protectable at the governance level.

The mechanism (§3–§4) responds to that gap. Cognitive sovereignty is the collective capacity to maintain independent, calibrated, and falsifiable judgment under the pressure conditions described, and its four dimensions specify its observability. The cognitive contradictor is the proposed protection mechanism: a system whose utility function is challenge, not validation, bound to four invariant principles so that the challenge is substantive and not performative.

Adoption (§6) translates that logic to the organizational level: mandatory challenge at the decision point, prospective and auditable recording of deliberation, role separation between whoever produces the analysis and whoever challenges it, verified calibration over time. The deliberative register — append-only and not retroactively modif-

iable — is the precondition that makes empirical calibration accessible and with it the improvement of the process (Tetlock & Gardner, 2015).

Research agenda. Three priorities emerge as necessary steps for the theoretical proposal to become an empirical contribution.

The first is the operationalization and validation of the construct dimensions. The operational definitions of §3.2 specify what each dimension would measure; translation into testable psychometric instruments — scales, observation protocols, coding procedures — is a not-yet-started program. Factorial structure, convergent validity, and discriminant validity with respect to adjacent constructs remain to be established empirically.

The second is the Phase 2 confirmatory test. As documented in *Calibrated Dissent* (Canepa, 2026, §6.5), the sealing-and-resolution protocol of the Osservatorio is designed to permit, on a larger, forward-sealed cohort, a pre-registered test of the calibration hypotheses (H-cal) and divergence-that-pays (H-pay). That phase will produce evidence on the forecasting quality of the contradictor — or, if the hypotheses do not hold, their falsification — and is the necessary step to transform the effectiveness claim from a theoretical proposal into an empirically grounded or rejected assertion.

The third is extension beyond M&A. The principles of the construct and the mechanism apply in principle to any high-stakes decision with low reversibility: public policy, institutional investment, healthcare governance under uncertainty. Verifying whether the erosion mechanisms (Bainbridge, 1983; Schwenk, 1990) manifest in the same forms — and whether the same principles counter them with the same effectiveness — requires independent research in those domains.

Conflict of Interest

The author is founder of CounterBrain / Coopera S.a.s. CounterBrain is cited in §4 exclusively as a reference implementation illustrating the achievability of the four design principles in an operational context; the mention does not imply any commercial recommendation nor a comparative evaluation of the adequacy of that implementation relative to alternatives. The design principles of the cognitive contradictor stated in §4 are presented as non-proprietary and reusable by anyone designing AI systems with an adversarial mandate. No external source of funding has influenced the content of this paper.

References

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