The Principle of Naturality: An Explainable Neuro-Symbolic Framework for Detecting Latent Convergence in Multi-Agent Inferential Structures*

Elio Javier Gogol Merletti
Information Systems Engineer; Court-Appointed Computer Forensics Expert
Higher-Education Professor of Programming, Computer Security, Software Modeling & Databases
3D.NET — Independent Software Development, Mendoza, Argentina
javier_gogol@hotmail.com ORCID: 0009-0004-4294-4816

Abstract

We present the Principle of Naturality (PN), an explainable neuro-symbolic framework that *certifies* latent convergences among multiple agents whose root premises are mutually exclusive—each convergence accompanied by a pairwise NLI certificate that is explicit about its backend (DeBERTa-v3-base, $\tau_{ent} = 0.55$, role-inversion verification) rather than an opaque score — so that improvements in the inference backend translate directly into improved certification. Each agent's stance is expanded into an actor-relative inferential tree ($P \to Q \to R \to \cdots$). A three-stage cascade certifies cross-actor compatibility: (i) sentence-embedding pre-filtering, (ii) bidirectional Natural Language Inference under pragmatic normalization, and (iii) role-inversion verification. A weighted path cost—Total Topological Friction (TTF)—selects convergence HUBs, and the selected consensus is characterized as a Pareto-efficient *utilitarian bargaining* solution. The certificate additionally induces a *canonical support geometry*: under four explicit axioms, every proposition receives a unique barycentric position stating which agents' certified material supports it; each HUB collapses to a single point, and in all four cases the utilitarian optimum is not the most equitable convergence available — the geometry prices the difference. The object PN emits—the convergence point, the per-agent causal paths, and the normative justification of its selection—is traceable end-to-end from the agents' axioms. We evaluate PN on four heterogeneous cases, including a fully fictional control, and benchmark it against geometric baselines (precision $\leq 7.2\%$) and a stage ablation; the cascade turns a cheap neural candidate set into a high-precision, auditable certificate, and the detected convergences are not an artifact of depth-driven semantic collapse. We are explicit about what remains illustrated rather than demonstrated—most centrally, expert validation of HUB quality, which is the only external reference, since PN's notion of convergence is itself model-relative.

Keywords: explainable AI, neuro-symbolic AI, natural language inference, sentence embeddings, graph theory, inferential trees, logic, multi-agent convergence, bargaining.

Contents

*An earlier version of this work received Second Place in the 2026 GCSP Prize for Transformative Futures in Peace and Security (Geneva Centre for Security Policy, Geneva, Switzerland), among 337 submissions from 95 countries, evaluated by an international jury including representatives of ETH Zurich, the University of Oxford, and the Swiss Federal Administration.

1 Introduction

1.1 Problem: conflicts stalled on premises

Conflicts between cognitively situated agents frequently stall not for lack of viable solutions but because mediation concentrates on incompatible *premises* [22]. Premises are anchored in values, history, and identity, and are the positions agents are least willing to modify. This work begins from an empirical observation: although agents defend incompatible premises, the *chains of consequences* they derive from those premises can converge at common points. In the Nile hydropolitical dispute, for instance, Egypt's historical-rights frame and Ethiopia's sovereignty frame are mutually exclusive at the premise level, yet both inferential chains reach shared operational consequents (regional stability, coordinated water management) that neither agent articulated as a common objective.

1.2 Observation: convergence lives in the consequences

PN operationalizes this observation. Each agent's position is expanded into an actor-relative inferential tree whose nodes are evaluable propositions and whose edges are the agent's own inferential commitments. Convergence is then sought not among premises but among the propositions these trees generate.

This choice of object has a cognitive motivation. Articulating *why* one holds a position—and then why that reason in turn holds, recursively—is an act of reasoning about one's own reasoning; the inferential tree is a formalization of this metacognitive articulation of justifications, not an arbitrary modeling convenience. It accords with accounts on which a central function of human reasoning is the production and evaluation of justifications [17], and it is what makes the central move available: by taking each agent's justificatory structure—rather than its stated position—as the unit of analysis, convergence can be sought *below* the level of what agents assert, in what their reasons commit them to. We make no stronger claim than this motivation requires. The tree operationalizes a single metacognitive act, the giving and chaining of reasons; it is not a model of cognition, nor of metacognitive monitoring, and we attach no cognitive interpretation to the terminal nodes of the tree beyond their being the points at which an agent's articulated justification, in a given run, stops.

1.3 The gap: detection is not explanation

Detecting that two agents *could* agree is of limited value to a mediator, an analyst, or an auditor unless one can also say *why*, *where*, and *at what cost to each party*. General-purpose explainable-AI techniques applied to a black-box agreement score produce post-hoc rationalizations detached from the agents' actual reasoning. In high-stakes settings—diplomacy, clinical ethics, legal adjudication—this gap is decisive: a mediator who cannot audit *why* two parties might agree cannot responsibly act on the claim, so an unexaminable agreement score is, for these purposes, scarcely better than no answer at all. PN is built so that the explanation is not added after the fact but is the primary object the method emits: a convergence point (HUB), the per-agent causal paths that reach it, and a normative justification of why that point is the rational consensus. This is the sense in which PN is an *explainable* neuro-symbolic framework, and the property this paper foregrounds.

1.4 Hypothesis

We state PN in two parts. **H1 (structural):** in conflicts between cognitively situated agents with legitimate interests, latent convergence points may exist within the joint deductive structure of the agents' positions even when their root premises are mutually exclusive. **H2 (operational):** when such convergences exist, they are detectable—and explainable—through logical-semantic analysis of each agent's consequence structure, without requiring premise modification. The absence of a global HUB is a valid diagnostic output, not a method failure.

1.5 Neuro-symbolic positioning and contributions

PN is neuro-symbolic in the compositional "Symbolic[Neural]" sense of Kautz [13], within the landscape surveyed by Garcez and Lamb [8]. The *neural* layer (sentence embeddings, cross-encoder NLI, LLM-driven tree expansion and role-inversion) acts strictly as an efficient *candidate generator*; the *symbolic* layer (proposition-level inferential commitments, a convergence graph, and a bargaining/social-choice selection rule) supplies an *auditable certificate* the neural layer cannot overrule. Unlike tightly integrated approaches such as Logic Tensor Networks [23] and DeepProbLog [16], where symbolic structure is learned or relaxed by gradient descent, PN keeps the symbolic certificate inviolable, which is what makes each convergence traceable.
The contributions of this paper are:

- a **three-stage cascade** validation architecture (embedding pre-filter; bidirectional averaged NLI with contradiction veto under pragmatic normalization; role-inversion verification);
- a **structured, auditable explanation** as the framework's output — the HUB, the per-agent causal paths, and the per-pair certificates — rather than an opaque agreement score;
- a **normative characterization of HUB selection** as a Pareto-efficient utilitarian bargaining solution over the agents' inferential costs (Total Topological Friction), distinguishing it from Nash and egalitarian solutions;
- a **canonical support geometry** for the certificate: an axiomatically unique (least-fixed-point) assignment of barycentric support coordinates to every proposition, under which each HUB occupies a single point and the equity of every convergence—who builds it vs. who pays to reach it—becomes a reported, measure-robust quantity;
- a **post-hoc Mapper visualization** of the unified semantic space, used to render the *shape* of disagreement and not as a constituent of detection; and
- an **empirical evaluation** over four heterogeneous cases including a fully fictional control that isolates framework behavior from language-model memorization, with a cosine-only baseline and a cascade stage ablation.

1.6 Document structure

The remainder of the paper is organized as follows. Section 2 walks a minimal two-agent example end to end. Section 3 formalizes the framework: inferential trees as subjective structures, the NLI cascade and the strict/lax validated cores, the convergence graph and HUB definitions, the characterization of HUB selection as a Pareto-efficient utilitarian bargaining solution, the divergence descriptors, and the post-hoc Mapper visualization. Section 4 details the five-phase pipeline. Section 5 evaluates PN on four heterogeneous cases and reports the cosine-only baseline and the

cascade stage ablation. Section 6 derives the canonical support geometry the certificate induces — the axiomatically unique barycentric position of every proposition — and its equity readout. Section 7 positions PN against related work, social choice, and argumentation, and states limitations; Section 8 concludes. Appendices collect notation, protocols, pseudocode, and the validated default configuration.

2 A Worked Example

Before the formal framework, we walk a deliberately minimal two-agent case end to end, so that every later definition can be read against a concrete referent. The example is small enough to verify by hand and exhibits the one phenomenon PN is built to detect and explain: two agents who disagree at the premise level nonetheless reach the *same* consequence through disjoint reasoning.

The conflict. A company must decide its strategy for the coming year under limited resources. This shared situation is the problem statement

$$\Pi = \text{“The company must decide its strategy for the next year with limited resources.”}$$

Two directors hold mutually exclusive positions: Actor A (pro-expansion) proposes expanding to new markets; Actor B (pro-consolidation) proposes consolidating current markets. Limited resources prevent executing both, so the *root premises cannot be reconciled.*

The inferential trees. Each director expands their own position into a short chain of consequences they themselves endorse (Tables 1 and 2). Neither tree mentions the other actor; each is built in the actor's own frame.

Table 1: Tree A (Expansion).

Level	Proposition
L_0 (root)	The company must expand to new markets.
L_1	The company invests resources in acquiring new customers.
L_2	Revenue from new customers increases.
L_3	The company improves its financial position.

Table 2: Tree B (Consolidation).

Level	Proposition
L_0 (root)	The company must consolidate current markets.
L_1	The company invests resources in retaining current customers.
L_2	Revenue from recurring customers increases.
L_3	The company's economic situation strengthens.

Why similarity alone is not enough. The two roots (L_0^A, L_0^B) are thematically close—both are one-sentence corporate strategies about “markets”—yet they are *contradictory*: expanding and consolidating under fixed resources cannot both hold. Cosine similarity over sentence embeddings

cannot separate "same topic, compatible" from "same topic, opposed"; this is precisely why PN routes every candidate pair through a logical cascade rather than trusting proximity.

The cascade. For each cross-actor pair the cascade applies: (i) an embedding pre-filter; (ii) bidirectional NLI under pragmatic normalization $\tilde{P}$; and (iii) role-inversion verification. The premise pair (L_0^A, L_0^B) is *rejected by contradiction*—the NLI head assigns high contradiction in both directions. The terminal pair (L_3^A, L_3^B), "the company improves its financial position" vs. "the company's economic situation strengthens," is *accepted*: bidirectional entailment is high and contradiction is low, so the cascade certifies the two propositions as *biconditionally equivalent under the cascade*—an explicit certificate relative to (NLI cross-encoder, τ_{ent}, role-inversion verification), not a claim of model-independent semantic identity—despite their different vocabulary.

The convergence and its cost. The single validated fusion induces a HUB at $L_3^A \approx L_3^B$. Each actor reaches it through its own chain, and the Total Topological Friction (Section 3) measures the inferential effort each pays. Table 3 gives the per-edge bicomposite weights $w(v) = \big(1-\mathrm{sim}(v, R_i)\big) + \big(1 - \mathrm{sim}(v, \Pi)\big)$ and the accumulated cost, computed with the multilingual S-BERT encoder [20]; summing each chain yields

$$\mathrm{TTF}(A) = 3.52, \quad \mathrm{TTF}(B) = 4.04, \quad \mathrm{TTF}_{\mathrm{sum}} = 7.56, \quad \mathrm{TTF}_{\mathrm{max}} = 4.04.$$

Table 3: Bicomposite edge weights and accumulated TTF for the worked example. R_i is the actor's root (L_0); v is the child proposition of each edge.

Actor	Edge	$\mathrm{sim}(v, R_i)$	$\mathrm{sim}(v, \Pi)$	$w(v)$	TTF (acc.)
A	$L_0 \to L_1$	0.709	0.417	0.875	0.875
A	$L_1 \to L_2$	0.527	0.131	1.342	2.217
A	$L_2 \to L_3$	0.368	0.331	1.301	3.517
B	$L_0 \to L_1$	0.536	0.391	1.072	1.072
B	$L_1 \to L_2$	0.275	0.144	1.581	2.653
B	$L_2 \to L_3$	0.368	0.242	1.390	4.044

What PN emits. The output is not the bare fact "the directors can agree." It is a *structured, auditable explanation*: the convergence point (financial improvement), the two disjoint causal paths that reach it, the per-pair certificate that licenses the equivalence, and the cost each party incurs. The agreement is an *inadvertent consensus*—both actors derive financial improvement without ever proposing it as a common goal—and the explanation is traceable end to end from the two root premises. This is the object every later section formalizes and the four empirical cases instantiate at scale.

3 Theoretical Framework

3.1 Inferential trees as subjective structures

Each agent's position is modeled as a directed inferential tree built in the agent's own frame of reference.

Definition 1 (Subjective inferential structure)**.** Let $T_{a_i} = (V_{a_i}, E_{a_i})$ be the tree of agent a_i. T_{a_i} is a *subjective inferential structure* if (a) each $v \in V_{a_i}$ is a proposition evaluable as true or false; (b) each edge $(u, v) \in E_{a_i}$ represents an inferential commitment $u \to v$ that a_i holds in their own frame; and (c) a_i assigns truth value T to all nodes of their own tree.

The edge operator is not material implication. The operator $\to$ must be read carefully, because the framework's whole logic depends on it. It is *not* the truth-functional material conditional. Within the agent-relative frame, $P \to Q$ denotes a *defeasible inferential commitment*: "given P, agent a_i commits to Q under bounded coherence," valid only inside that agent's reasoning, not as a globally true/false proposition. Three consequences follow. First, the framework neither requires nor produces a global truth assignment over $V = \bigcup_i V_{a_i}$; truth is local to each agent. Second, the same edge may be endorsed by one agent and absent from another's tree without contradiction. Third, because the commitment is defeasible, deeper consequences are entertained *relative* to the chain that produced them—exactly the "if circumstances develop this way" reading used in the worked example (Section 2), where "the company invests in new customers $\to$ revenue increases" is the director's own commitment, not a universal law. Each node carries an identity tuple $(\mathsf{actor\_id}, \mathsf{level}, \mathsf{index})$ for full traceability; each tree is rooted at the agent's declared position $R_i \in V_{a_i}$; and a shared *problem statement* Π anchors conflict relevance in the edge weight of Section 3.3.

Branches as alternative consequence chains. A reasoning tree is built by best-first expansion of consequences, so it naturally contains branches that are mutually exclusive: an agent may explore "if circumstances develop one way, p" on one branch and "if they develop another way, $\neg p$" on a sibling branch. We treat each root-to-leaf branch as a *hypothetical consequence chain* from the agent's stance under a particular development of circumstances, and the whole tree as the family of such chains the agent considers. Two opposed branches assert p under one developmental hypothesis and $\neg p$ under another; we read this as the agent mapping alternative possibilities conditionally, not as the agent asserting $p \wedge \neg p$ simultaneously. No branch holds a contradiction; cross-branch opposition is inferential exploration of disjoint hypothetical conditions.[1]

3.2 Semantic equivalence and the fusion relation

Vectorization. Propositions are mapped to vectors with a multilingual Sentence-BERT encoder [20], $\varphi : U \to \mathbb{R}^{768}$; semantic proximity is the cosine similarity $\mathrm{sim}(P, Q) = \varphi(P)\varphi(Q)/(\|\varphi(P)\|\|\varphi(Q)\|)$. Cosine detects *thematic* association but not *inferential* compatibility: two propositions can be thematically near yet contradictory (the agency-inversion case below), which is why proximity only *pre-filters* candidates and a logical cascade decides.

NLI as a map onto the probability simplex. The cross-encoder NLI model [10], fine-tuned on SNLI/MNLI [3, 28], takes an *ordered pair of texts* (not the $\mathbb{R}^{768}$ embeddings, which serve only the pre-filter) and returns three logits whose softmax is a point on the 2-simplex $\Delta^2 = \{(e, n, c) : e, n, c \geq 0,\ e+n+c = 1\}$—a probability distribution over ENTAILMENT, NEUTRAL, CONTRADICTION, geometrically an equilateral triangle with those three classes at its vertices and barycenter $(\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$ the point of maximal uncertainty. Fusion is then a *region* of this simplex: high entailment, far from the contradiction vertex.

[1] This is an operational interpretation of the tree expansion under the bounded best-first procedure of Section 4; we do not invoke any formal modal or possible-worlds semantics, since the branches are heuristically truncated and not closed under deduction.

Pragmatic normalization redirects attention. Each agent declares a formal name n_i and a short alias α_i. The operator $\tilde{P}$ replaces every occurrence of α_i (and its possessive) by a generic token ACTOR before inference. The justification is mechanistic: a proper noun is a high-salience token that a transformer's self-attention concentrates on, carrying learned name-specific associations (e.g. the rivalry priors attached to "Egypt" or "Ethiopia"); replacing it with ACTOR *redirects attention* from the name toward the causal/agency architecture of the statement. We observed this empirically: a holistic judgment of the same pairs shifts markedly when names are normalized away (and $\tilde{P}$ is decisive on the GERD case), confirming that without it the classifier reacts to identities rather than to inferential content. Note $\tilde{P}$ normalizes only the *owner* alias; third actors mentioned remain, since mentioning a third party is substantive content, not speaker identity.

The fusion relation F. For two candidate nodes (P, Q) the bidirectional criterion is

$$F(P,Q) = \mathsf{True} \iff \underbrace{\mathrm{AVG}\big(\mathrm{NLI}(\tilde{P}\to\tilde{Q})_e, \mathrm{NLI}(\tilde{Q}\to\tilde{P})_e\big) \geq \tau_{ent}}_{\text{bidirectional entailment}} \wedge \underbrace{\mathrm{MAX}\big(\cdots_c\big) < \tau_{contr}}_{\text{contradiction veto}},$$

with $\tau_{ent} = 0.55$, $\tau_{contr} = 0.30$. We call F the *fusion relation*: the set of cross-actor pairs *certified as biconditionally equivalent under the cascade*—a phrasing we use deliberately to mark that F is not a claim about model-independent semantic identity but a relation defined by, and dependent on, the specific cascade (NLI cross-encoder, τ_{ent}, contradiction veto, role-inversion LLM). The *bidirectional* requirement has a precise logical reading: the cascade's equivalence is *mutual entailment*, $P \models Q \wedge Q \models P$, i.e. the biconditional $P \leftrightarrow Q$. A single direction would give only implication (Q a consequence of P), not equivalence; evaluating both directions operationalizes the biconditional and mitigates the agency-polarity asymmetries single-direction NLI misses ("ACTOR retains authority over flow" vs. its agency-inverted variant).

The AVG aggregate alone, however, can be cleared by a strongly *asymmetric* pair: directional scores $(0.95, 0.20)$ average to $0.575 > \tau_{ent}$ while constituting unidirectional implication rather than biconditional equivalence. The cascade closes this gap with its guards rather than with the aggregate: the contradiction veto is evaluated in *both* directions, and the role-inversion stage (iii below)—whose declared function is precisely to reject pairs whose agency or polarity does not survive inspection—removes the asymmetric survivors (14–22% of the pairs the NLI stage accepts; Table 5). As a robustness check we additionally *re-certified* the entire corpus under the strictly stronger MIN aggregate (each direction individually $\geq \tau_{ent}$), guards unchanged: 74–86% of F survives and the utilitarian HUB selection is essentially invariant (Section 7.3, *aggregation-criterion robustness*). The AVG criterion with both guards is the operating point of every result in this paper; MIN is reported as its stress test.

The full three-stage cascade wraps this criterion in two guards. For u, v with $\mathrm{actor}(u) \neq \mathrm{actor}(v)$, $F(u,v)$ holds iff: (i) the embedding pre-filter $\mathrm{sim} \geq \tau_{pre} = 0.35$; (ii) the bidirectional NLI criterion above under $\tilde{P}$; and (iii) an *LLM role-inversion verification* (Section 4) that catches pairs which pass NLI yet have inverted agency or swapped third-party references. F is completed by reflexivity and symmetry, and is empirically *not* transitive: there exist $F(u,v) \wedge F(v,w)$ with $\neg F(u,w)$ — in the four corpora, 92–100% of length-two F-chains fail to close (GERD 92.0%, Selene 93.8%, Lia Lee and Bipolar I 100%; artifact `transitividad_F.py`).

From F to classes via Union–Find. To find multilateral convergences we treat F as a graph (nodes = propositions, edges = fusions) and compute its *connected components* with the disjoint-set Union–Find algorithm [7, 25], near-linear at $O(m\,\alpha(n))$: every two propositions transitively linked by F are grouped into one *candidate class* (the transitive–symmetric closure $\approx_F$). A class is *not*

yet a HUB: because F is non-transitive, a component can chain ($u \sim v \sim w$ without $u \sim w$). Within each class we therefore read off two validated cores.

Definition 2 (Strict and lax validated cores)**.** Let C be a candidate class with present actors $A(C) = \{a_1, \ldots, a_k\}$. The *strict core* is the union of nodes in some closed multipartite k-clique under F: $\text{core}_{strict}(C) = \bigcup\{K = \{n_1, \ldots, n_k\} : n_i \in V_{a_i}, F(n_i, n_j)\ \forall i \neq j\}$. The *lax core* is the set of nodes with complete cross-actor coverage: $\text{core}_{lax}(C) = \{v \in C : \forall a' \neq \text{actor}(v), \exists u \in C \cap V_{a'} \text{ with } F(v, u)\}$, with $\text{core}_{strict} \subseteq \text{core}_{lax} \subseteq C$.

The certificate, and why it needs no transitivity. By *certificate* we mean precisely this object: for an accepted convergence, the set of per-pair F-decisions (and the clique they form) that license it, each traceable to a specific NLI/role-inversion judgment. A *strict* clique is the strongest certificate: within it the equivalence is *complete*—every one of the k representatives is pairwise F-validated with every other, *simultaneously*, not by chaining. Crucially, the clique does *not* borrow a transitivity F does not have: it requires each pair to be *directly* in F. (The induced relation restricted to a clique is trivially transitively closed, but only because completeness makes it so, achieved by direct validation.) This is exactly why no tolerance- or rough-set machinery is needed *for multilateral certification*: connected components give candidate classes, and closed cliques give the certified multilateral equivalence, with lax coverage as the weaker fallback when no clique exists.

3.3 Graph theory and HUB detection

The validated relation F is assembled into a unified directed graph with three layers: the *causal graph* G_C (the per-actor DAGs, acyclic by construction), the *semantic compatibility graph* G_F (cross-actor validated pairs), and the *operational closure graph* G_O induced by the transitive–symmetric closure $\approx_F$. Only G_C is guaranteed acyclic.

Definition 3 (Strict and lax HUB)**.** Let $[n]$ be an operational grouping with actor set $A_{[n]}$. A *strict HUB* $h_s = (K, [n])$ is a closed multipartite clique $K \subseteq \text{core}_{strict}([n])$ with $|K \cap V_{a_i}| \geq 1$ for every $a_i \in A_{[n]}$ and $(u, v) \in F$ for every cross-actor pair in K. A *lax HUB* $h_\ell = (R, [n])$ is a representative set $R \subseteq \text{core}_{lax}([n])$, one node per actor, such that each r_{a_i} has direct F-validation with at least one node of every other actor (the pairs (r_{a_i}, r_{a_j}) themselves need not be in F). Lax HUBs are emitted only when no strict HUB exists for $[n]$.

A HUB is a point at which every actor's inferential chain reaches a common, pairwise-certified consequence. Each actor arrives through its own chain in T_{a_i} and never adopts external logic; the explanation is the HUB together with these chains.

Deduplication of strict HUBs. Definition 3 admits multiple distinct strict HUBs over a single convergence region. When the UF closure of a HUB contains more than one node per actor, every combination of one node per actor that is pairwise F-connected forms a closed multipartite clique and is therefore an independent strict HUB by Definition 3, even though all such HUBs share the same closure. For example, in the largest equivalence class of Lia Lee, 64 strict HUBs share an identical UF-closure of 39 nodes and differ only in their two-element representative cores. We audited this directly: across the four cases, every pair of strict HUBs with identical closure has *distinct* representative cores (0 pairs of literally identical HUBs in 3,375 pairs of closure-identical HUBs), so the multiplicity is a property of the clique structure, not an artifact of the implementation.

For reporting and selection we therefore group strict HUBs into equivalence classes by UF-closure (Jaccard ≥ 0.5 on clase_nodos; the conservative threshold 0.8 yields the same partition since

closures are either identical or disjoint in practice), and we report counts of equivalence classes rather than of clique instances. All HUB counts in Section 5 refer to these equivalence classes. The internal multiplicity within a class is itself informative—it reflects the density of F-connectivity within the convergence region—and is preserved in the `duplicate_of` field of `hubs.json` for downstream uses that need clique-level granularity.

Definition 4 (Bicomposite edge weight)**.** For an edge (u, v) in tree T_{a_i} with root R_i and problem statement Π,

$$w(u,v) = \big(1 - \mathrm{sim}(\varphi(v), \varphi(R_i))\big) + \big(1 - \mathrm{sim}(\varphi(v), \varphi(\Pi))\big).$$

The first term measures *identity drift* from the actor's root; the second measures *conflict relevance* to the dispute. Both terms are non-negative, so Dijkstra's algorithm [4] is correct on T_{a_i}.

Definition 5 (Total Topological Friction)**.** For each actor $a_i \in A_{[n]}$, $\mathrm{TTF}(a_i, h) = \mathrm{dist}_{T_{a_i}}(R_i, r_{a_i})$, the accumulated bicomposite weight from the actor's root to its representative in the HUB. Aggregate metrics: $\mathrm{TTF}_{\mathrm{sum}}(h) = \sum_i \mathrm{TTF}(a_i, h)$ and $\mathrm{TTF}_{\mathrm{max}}(h) = \max_i \mathrm{TTF}(a_i, h)$.

Definition 6 (Lexicographic HUB selection)**.** Let $\sigma(h) = 0$ if h is strict and $\sigma(h) = 1$ if lax—a binary *certificate-strength* indicator, not to be confused with the NLI thresholds $\tau_{ent}, \tau_{contr}, \tau_{pre}$. The selected HUB is $h^* = \arg\min_{h \in H} \big(\sigma(h),\ \mathrm{TTF}_{\mathrm{sum}}(h),\ \mathrm{TTF}_{\mathrm{max}}(h)\big)$, ordering certificate strength above aggregate cost above worst-case cost.

Inadvertent consensus. A HUB is an *inadvertent consensus*: the actors begin from incompatible premises, follow divergent chains, and arrive at a shared consequence none proposed as a common goal. The worked example (Section 2) is the minimal instance; the empirical cases (Section 5) exhibit it at scale.

Scope. A HUB certifies *logical* reachability of a common consequence through validated actor-relative paths, not its *social* acceptability; the latter requires subsequent human negotiation. PN is a structured discovery procedure, not an automated mediator.

3.3.1 HUB selection as a bargaining solution

The lexicographic rule of Definition 6 is more than a tie-break: it is a normative choice over which agreement to return, and it admits a precise characterization in cooperative bargaining terms. Fix an operational grouping with candidate HUBs H. Each HUB h induces a cost vector $c(h) = (\mathrm{TTF}(a_1, h), \ldots, \mathrm{TTF}(a_k, h))$: the inferential effort each agent pays to reach it. Treating $-c_i$ as agent i's utility, the candidate HUBs form a discrete feasible set of agreements, and Π (no agreement) is the disagreement point.

PN's primary criterion, $\arg\min_h \mathrm{TTF}_{\mathrm{sum}}(h) = \arg\min_h \sum_i c_i(h)$, is the *utilitarian* bargaining rule. This is distinct from the *Nash* solution [18] ($\arg\max_h \prod_i (d_i - c_i(h))$, which favors balanced gains) and from the *egalitarian* solution [12] ($\arg\min_h \max_i c_i(h)$, which PN uses only as a tie-break through $\mathrm{TTF}_{\mathrm{max}}$).

Remark 1 (Pareto efficiency of the utilitarian HUB)**.** The HUB selected by minimizing $\mathrm{TTF}_{\mathrm{sum}}$ is Pareto-efficient over the candidate set: no other candidate HUB h' satisfies $\mathrm{TTF}(a_i, h') \leq \mathrm{TTF}(a_i, h^*)$ for all i with strict inequality for some. If such h' existed, summing the inequalities over i would give $\mathrm{TTF}_{\mathrm{sum}}(h') < \mathrm{TTF}_{\mathrm{sum}}(h^*)$, contradicting the minimality of h^*. This is the standard observation that any utilitarian (sum-minimizing) selection over a finite feasible set is Pareto-efficient; we record it for the reader who wants the bridge from the lexicographic rule to

social-choice vocabulary, not as a substantive contribution. The substantive content of this subsection lies in the empirical placement that follows.

Empirical placement. Across the four corpus cases, the TTF_{sum}-selected HUB is Pareto-efficient in 4/4 cases (as Proposition 1 guarantees) and coincides with the Nash bargaining solution in 3/4. The exception is GERD: PN's selected HUB has cost vector $(8.01, 5.78, 3.82)$—a low total but an *unequal* split (Sudan pays little, Egypt much), whereas the Nash solution prefers the more balanced $(7.89, 5.73, 6.73)$. This is the textbook divergence between utilitarian and Nash bargaining: PN minimizes total inferential effort even when one party bears most of it, and does not enforce an equitable distribution. This is a transparent, debatable normative commitment—precisely the kind a mediator must inspect—and stating it explicitly is part of what makes PN's output an explanation rather than a verdict.[2]

3.4 Semantic divergence analysis

Beyond the convergence points, PN summarizes how each actor's reasoning moves relative to the dispute.

Definition 7 (Semantic divergence gradient)**.** For an edge $(u, v) \in E_i$ in tree T_{a_i}, the divergence gradient with respect to the problem Π is $\Delta\text{sim}_\Pi(u, v) = \text{sim}(\varphi(v), \varphi(\Pi)) - \text{sim}(\varphi(u), \varphi(\Pi))$, positive when a step moves toward the problem and negative when it moves away.

Definition 8 (Actor divergence coefficient)**.** For an actor a_i with tree $T_{a_i} = (V_i, E_i)$, $D(a_i) = |\{e \in E_i : \Delta\text{sim}_\Pi(e) < 0\}|/|E_i|$. A value $D(a_i) > 0.5$ indicates a tree whose derivation steps predominantly move away from Π.

Together, Δsim_Π diagnoses the dynamics of each derivation step and $D(a_i)$ summarizes the divergence propensity of an actor's tree; both are diagnostic complements to TTF (Definition 5), which integrates identity drift and conflict relevance along the bicomposite-weighted path. Reported per actor and per case (Section 5.4), the pair $(D(a_i), \text{TTF}(a_i, h^*))$ forms a compact *divergence profile*: it distinguishes an actor that drifts away from the dispute yet still reaches the consensus cheaply from one whose every step toward it is costly—information a mediator needs but a single agreement score hides.

3.5 Mapper visualization of disagreement (post-hoc)

The detection pipeline (Sections 3.2–3.3) is self-contained: HUBs, validated cores, and TTF rankings do not depend on any visualization. As a separate, post-hoc layer, PN renders the *shape* of a disagreement using a Mapper skeleton [24, 21] of the unified semantic space (Figure 1). This is a presentation device for the topology already certified by F, not a constituent of detection.

3.6 A descriptive probe for external mediations

The machinery built for HUB selection supports a second, strictly descriptive use, which we report as an application rather than a core claim. A HUB is an *endogenous* convergence—a point the

[2] Cost vectors are read from the per-HUB `distancias` field of the case outputs; the comparison is computed in the released artifact (`nash_bargain.py`). The Nash coincidence is sensitive to the choice of disagreement point, so we claim only the utilitarian characterization and its Pareto efficiency, not identity with Nash. The Nash-preferred alternative ($\text{TTF}_{\text{sum}} = 20.36$) is also the one GERD HUB that does not survive the MIN re-certification of Section 7.3, so the utilitarian-vs-Nash contrast is a property of the AVG operating point.

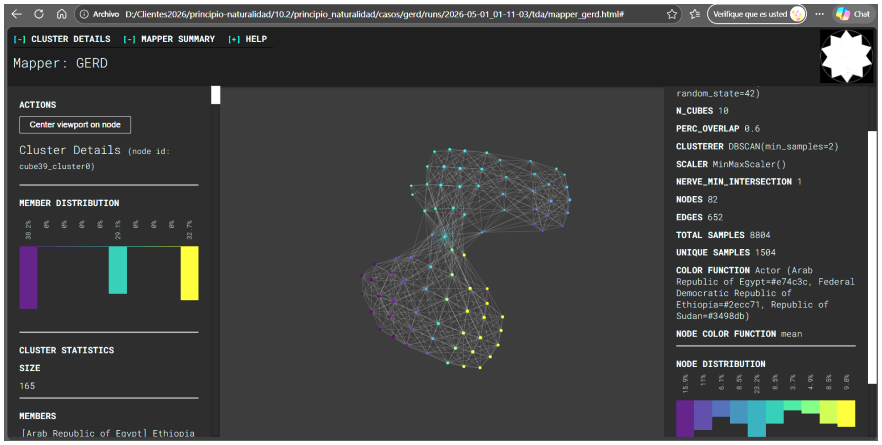

Figure 1: Mapper skeleton of the unified semantic space (GERD): a visualization of the *shape* of a disagreement. Post-hoc; not part of the detection pipeline.

agents already reach. A mediator, however, arrives with *exogenous* proposals and needs to know, before tabling one, which parties' reasoning can already accommodate it and who must still move. PN answers this with the same objects (the trees, the cascade, and TTF) and without leaving the descriptive regime.

The frozen tree is the epistemic guarantee. An external mediation P_m is treated as one additional proposition and tested against each agent's *existing* tree. We do *not* re-expand the tree toward P_m: each tree was generated to justify its agent's stance, by best-first expansion of that agent's own consequences, neutral to any mediation. Re-generating toward a target would let the inference engine manufacture a path to almost any proposition—convergence by construction—and would destroy descriptive validity. Keeping the tree frozen is therefore not a limitation but the guarantee: a detected support is genuine (the agent's own neutral reasoning already reached it), and a non-detection is meaningful (best-first prioritizes where the agent's reasoning actually goes, so failing to reach P_m is informative about that agent's justificatory structure, not a sampling accident).

Two quantities per agent. For agent a_i we report (i) the *matching node*—the proposition $v \in T_{a_i}$ that supports P_m, an auditable trace—and (ii) the *support* level $\mathrm{supp}(a_i, P_m) = \max_v \mathrm{ent}(v \to \tilde{P}_m) \in [0, 1]$, the strongest non-contradicting entailment of P_m by the agent's reasoning under pragmatic normalization. We use one-directional entailment, not the biconditional fusion relation F of Section 3.2: a well-formed mediation is new content, not a restatement of an existing node, so equivalence is the wrong test (it returns no support everywhere). To control the false positives this relaxation admits—an entailment can fire on shared vocabulary while the agent's stance is in fact opposed—we apply the full cascade discipline to acceptance: a bidirectional contradiction veto and the role-inversion LLM verification of Section 4. The incorporation cost is the friction to the supporting node, $\mathrm{cost}(a_i, P_m) = \min_{v:\, v \models \tilde{P}_m} \mathrm{TTF}(R_i, v)$.

Convergence profile and the mediator's maximin loop. The vector $(\mathrm{supp}(a_1, P_m), \ldots, \mathrm{supp}(a_k, P_m))$ is the *convergence profile* of P_m: who is close, who is far. Its floor $\tau^*(P_m) = \min_i \mathrm{supp}(a_i, P_m)$ is the threshold at which all agents support P_m, and the *binding agent* $\arg\min_i$ is the party that must move. A mediator then iterates over the *proposal* (never over the trees), rewording P_m to raise

the floor—maximizing $\min_i \text{supp}(a_i, P_m)$—until either $\tau^* \geq \tau_{ent}$ (a mediation supported by all) or no rewording lifts the binding agent. This is exactly the *egalitarian* (maximin) criterion, the third bargaining rule of Section 3.3.1, now applied to external proposals with the agents' positions held fixed. The reported gap $= \max(0, \tau_{ent} - \tau^*)$ tells the mediator how far the binding agent still is, and the support dispersion $\max_i \text{supp} - \min_i \text{supp}$ how unbalanced the proposal is.

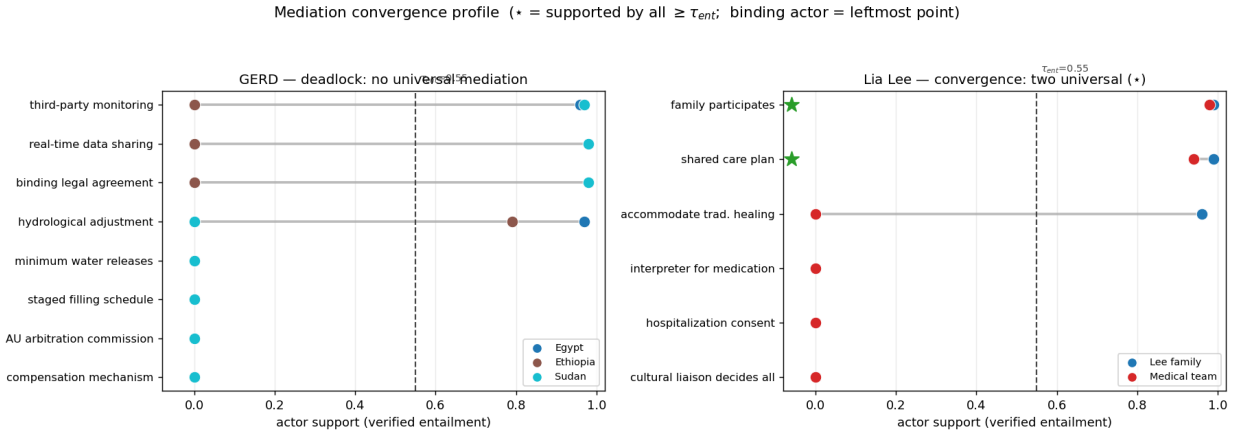

Figure 2: Mediation convergence profiles (verified support, post role-inversion). Each row is a candidate atomic mediation; each dot an agent's support; the gray segment its dispersion; ⋆ marks a mediation supported by all agents at τ_{ent}; the binding agent is the leftmost dot. GERD (left): every proposal leaves at least one agent at the floor—a structurally visible deadlock with no universally pre-supported atomic mediation. Lia Lee (right): two proposals are supported by both parties, and one (*accommodate traditional healing*) is supported by the family but not the team, which identifies precisely where a mediator would work.

What this shows, and its limits. Across the four cases the probe behaves as a descriptive aid, not a solver. It identifies, per proposal, which parties' reasoning already supports it and who is the hold-out, and these hold-outs are substantively faithful (Ethiopia on external oversight in GERD; the clinical team on accommodation in Lia Lee). It does *not* claim to discover universal solutions: under the full cascade, atomic external proposals are rarely pre-supported by all parties—which is consistent with, rather than contrary to, the deadlock these conflicts exhibit. Three limits bound the reading. The support signal inherits the NLI model's fragility, so the *matching node*, not only its score, must be inspected (the same caveat as the contradiction label, Section 7.3). A negative verdict is relative to the bounded tree: it states that the agent's *generated* justificatory reasoning does not reach P_m, not that the agent would reject it. And "incompatible" is always "no supporting wording was found," never a proof of impossibility, since the space of rewordings cannot be exhausted. Within these bounds, the probe is an explainable instrument for mediation design: it returns the trace, the hold-out, and the gap to close, while the creative work of reformulation remains with the human mediator.

4 Method

The pipeline runs in five phases and is independent of how the inferential trees are populated: by an LLM, by the actors themselves, or by external auditors. The language model is a modular, interchangeable component, never the source of the convergence claims.

Phase 1 — Tree construction. For each actor the tree T_{a_i} is built by best-first search with priority given by the accumulated bicomposite weight $W(v)$ (Definition 4). The inference engine receives the complete branch and returns candidate consequences (atomic propositions of 4–12 words, third person, mediation prohibited). A double-blind protocol prevents the engine from seeing other actors' positions, blocking artificial mediation bias. Termination is guaranteed by causal exhaustion, a depth limit ($d_{\max} = 10$), or a node budget ($n_{\max} = 500$ per actor). Two filters shape the expansion: an *admission gate* discards a candidate whose conflict relevance falls below $\tau_{\text{adm}} = 0.15$ (i.e. $\text{sim}(v, \Pi) < \tau_{\text{adm}}$), spending the node budget on dispute-relevant consequences; and a *cross-branch redundancy* filter discards a candidate whose cosine similarity to an existing tree node exceeds 0.90. These similarities, and the root embedding used for the identity-drift term, are computed on actor-name–neutralized text (the $\tilde{P}$ normalization of Section 3.2) to avoid lexical bias from proper nouns.

Phase 2 — Cascade validation and validated cores. Phase 2a applies the three-stage cascade (Section 3.2) to every cross-actor pair—(i) the embedding pre-filter, (ii) bidirectional NLI under pragmatic normalization, and (iii) an *LLM role-inversion verification* that rejects pairs whose agency, causality, or third-party references do not survive inspection—producing the set F of directly validated fusions. Stage (iii) is the only point at which an LLM enters validation, and it can only *reject*, never create, a fusion. Phase 2b runs Union–Find over F to obtain operational groupings (connected components). Phase 2c computes, for each non-trivial grouping, the strict cliques and the lax core (Definition 2).

Phase 3 — Graph assembly. The unified graph is assembled from the causal layer G_C (acyclic per-actor DAGs), the compatibility layer G_F, and the operational closure G_O.

Phase 4 — HUB search. For each grouping with universal actor coverage, PN emits one strict HUB per closed multipartite clique; if no strict clique exists but the lax core is non-empty, exactly one lax HUB is emitted. All HUBs are ordered by $(\sigma, \text{TTF}_{\text{sum}}, \text{TTF}_{\max})$ and the optimum is selected (Definition 6).

Phase 5 — Solution synthesis (optional). A natural-language rendering of the selected HUB and the two actor paths can be produced for human readers; it is presentational and plays no role in detection.

Complexity and configuration. Phase 1 is dominated by inference-engine latency at $O(kL)$ calls; Phase 2a is $O(n^2/B)$ for batched NLI, with the embedding pre-filter substantially reducing the evaluated population (e.g. 54% pair elimination in GERD); Phase 2b is $O(m\,\alpha(n))$ via Union–Find; Phase 4 reuses cached Dijkstra distances. The pre-filter trades completeness for tractability: a pair below τ_{pre} is never sent to NLI, so a true equivalence with low embedding similarity could be pruned before validation. This is a deliberate recall-for-cost trade, not a soundness gap—every emitted fusion remains fully certified by stages (ii)–(iii); only recall, never precision, is placed at risk. The configuration is common to all four cases: embeddings `paraphrase-multilingual-mpnet-base-v2`; NLI `cross-encoder/nli-deberta-v3-base`; LLM expansion `claude-sonnet` at temperature 0; thresholds $\tau_{pre} = 0.35$, $\tau_{ent} = 0.55$, $\tau_{contr} = 0.30$; $d_{\max} = 10$, $n_{\max} = 500$. A full configuration listing and the reproducible artifact are provided with the paper.

5 Empirical Validation

We evaluate PN on four heterogeneous cases spanning two actor cardinalities and three structural conflict types: a fully fictional control (Selene, three actors), a real geopolitical trilateral dispute (GERD, three actors), an intrapersonal conflict (Bipolar I, two cognitive states), and an intercultural medical-ethics conflict (Lia Lee, two actors). All four use the common configuration of Section 4. For each case we show the convergence graph G_F and the Mapper visualization of the unified semantic space.

5.1 Selene: a fully fictional control

Selene is a synthetic three-actor scenario—the space agencies of Aurelia, Borealis, and Cerulea sharing a lunar habitat with life-support insufficient for the combined crew. Nations, setting, and conflict are invented and have no precedent in any training corpus, so detected convergences cannot be attributed to language-model memorization of prior outcomes.

Veto-versus-substantive contrast (ablation). An initial formulation (v1) of the three stances used categorical clauses ("refuses any framework that…", "rejects any framework that…") that close the logical space of implications by construction. Under v1, PN produced *zero* strict and zero lax HUBs. A second formulation (v2) rewrote *only* the closing clauses, replacing absolute vetoes with red lines anchored in concrete functional reasons (technical certification, operational responsibility, crew safety) while preserving the substantive asymmetries and the absence of any conciliatory disposition. Under v2, PN identified **two unique strict HUBs** (8 clique instances) **and one lax HUB** under the biconditional cascade criterion of Section 3.2, with IPVR = 1.0 over the representatives of the top HUB—the *Internal Pairwise Validation Rate* is the fraction of cross-actor representative pairs directly certified by F, so 1.0 marks a closed clique. Detection behavior is driven by the formal architecture, not by prior knowledge in the LLM—PN fabricates no convergence when the implication space is closed by construction, and finds convergences when it admits a reachable common point. The control makes the division of labor vivid: the LLM *proposes* and PN *disposes*. The same generator supplies candidate consequences for both formulations, but whether they certify into a convergence is decided by structure alone—which is why v1 yields nothing and v2 yields a distinct strict HUB from the very same neural component.

5.2 GERD: a hydropolitical trilateral dispute

The Nile dispute over the Grand Ethiopian Renaissance Dam involves three state actors over a shared transboundary resource [19]. Egypt invokes the "no significant harm" principle and historical rights; Ethiopia invokes "equitable and reasonable utilization" and rejects the colonial-era treaties; Sudan conditions acceptance on a robust real-time technical data-exchange mechanism.

HUBs. PN produced **three unique strict HUBs** under the biconditional cascade criterion: hydrological unpredictability ($\text{TTF}_{\text{sum}} = 17.61$), a Sudanese diplomatic-reorientation convergence ($\text{TTF}_{\text{sum}} = 20.36$—the balanced alternative the Nash solution prefers in Section 3.3.1), and African-Union mediation ($\text{TTF}_{\text{sum}} = 21.39$). In the utilitarian-selected hydrological HUB, Sudan emerges as the structural pivot, reaching the convergence with the shortest inferential path (distance 3.82, shorter than Ethiopia's 5.78 and Egypt's 8.01).

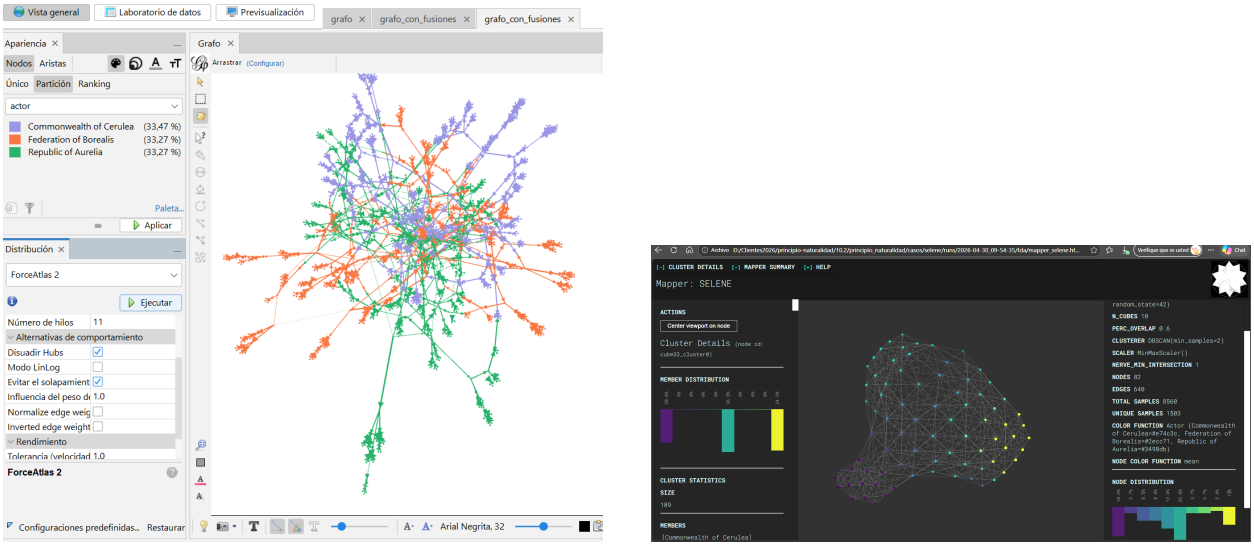

Figure 3: Selene. Left: convergence graph G_F with fusions. Right: Mapper skeleton of the unified semantic space.

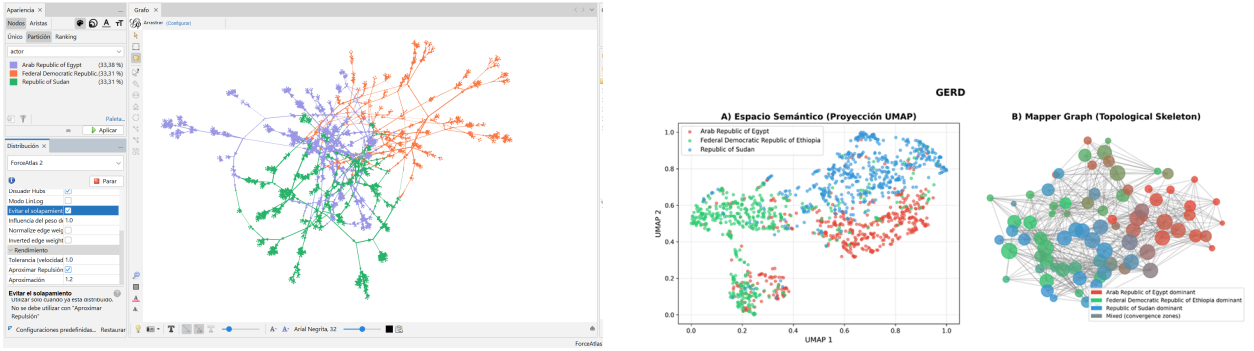

Figure 4: GERD. Left: convergence graph G_F. Right: Mapper skeleton; Egypt and Ethiopia occupy the most divergent regions, with Sudan as structural pivot.

5.3 Bipersonal cases

Bipolar I (Jamison). An intrapersonal conflict between two antagonistic cognitive states of one individual (manic, depressive), each assigned an exclusive, non-overlapping bibliography [9, 11]. PN produced **11 unique strict HUBs** (38 clique instances) under the biconditional cascade criterion of Section 3.2, concentrated around structural attractors such as decreased insight and isolation from the treatment system (Figure 5).

Intercultural medical ethics (Lia Lee). A bicultural bioethical case (Hmong spiritual etiology vs. Western clinical protocol) [6] produced **37 unique strict HUBs** under the biconditional cascade criterion of Section 3.2, concentrated around four dominant convergence classes (status-epilepticus risk, monitoring escalation, communication breakdown, redirection of family resources; Figure 6).

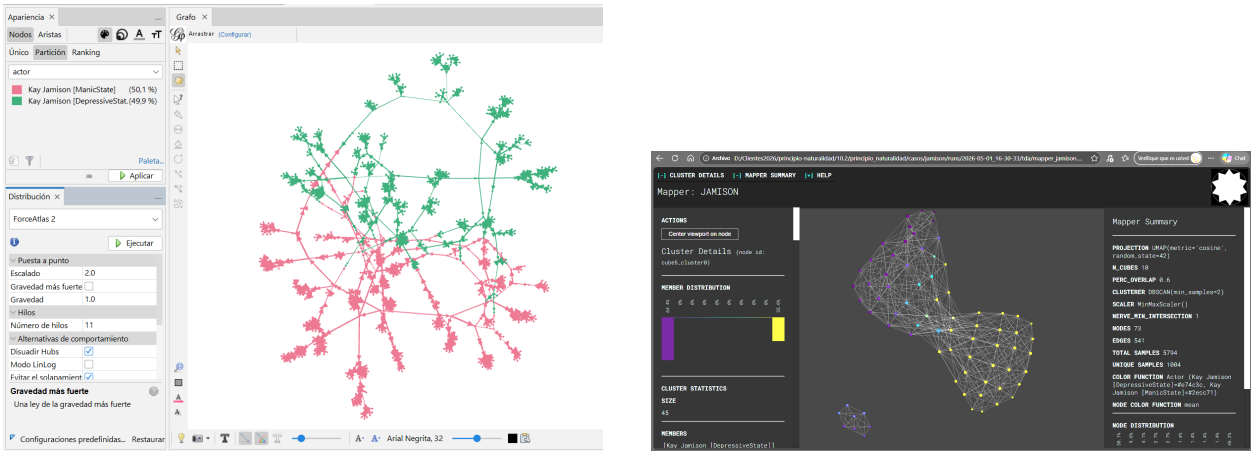

Figure 5: Bipolar I (Jamison). Left: convergence graph G_F between the manic and depressive cognitive states. Right: Mapper skeleton.

Figure 6: Lia Lee. Left: convergence graph G_F between the Lee family and the medical team. Right: Mapper skeleton.

5.4 Cross-case summary and the bargaining placement

Across the corpus, the $\mathrm{TTF}_{\mathrm{sum}}$-selected HUB is Pareto-efficient in 4/4 cases (Proposition 1) and coincides with the Nash bargaining solution in 3/4; the exception (GERD) is the utilitarian-vs-Nash divergence analyzed in Section 3.3.1—PN accepts the lower-total but less equal split $(8.01, 5.78, 3.82)$ where Nash would prefer the balanced $(7.89, 5.73, 6.73)$.

Per-HUB inferential profile. Beyond the per-actor cost $\mathrm{TTF}(a_i, h)$, each HUB admits a complementary, mass-side description that the bargaining rule itself does not use but a mediator can read. Let $D = \bigcup_i T_{a_i} \cup F$ be the directed graph induced by the actor trees (parent→child) together with the bidirectional fusion edges. For a HUB h with UF-closure $C(h)$ we define the *inferential mass*

$$m(h) \;=\; \sum_{a_i \in A_{[n]}} \; \sum_{x \in C(h)} \; \sum_{\pi:\, R_i \rightsquigarrow x \text{ in } D} \; \prod_{e \in \pi} w(e),$$

a cap-bounded enumeration over the actors' weighted paths into the closure, and the per-actor contribution $a_i(h)$ as i's share of $m(h)$ (so $\sum_i a_i(h) = 1$). The pair $\big(m(h),\, a(h)\big)$ summarizes "how much the agents construct h together" and "who builds it", as opposed to TTF which summarizes "who pays how much". From $a(h)$ we classify each HUB as *equilibrated* when $\max_i a_i(h) < 1/k{+}0.10$ (balanced contribution), *dominated* when $\max_i a_i(h) > 0.70$ (one actor carries the convergence), and *asymmetric* otherwise. Across the four cases the utilitarian-selected HUB ($\arg\min \mathrm{TTF}_{\mathrm{sum}}$) and the maximum-mass HUB coincide in three cases (GERD, Bipolar I, Selene) and diverge in Lia Lee, where the cost-minimal HUB and the mass-maximal HUB are distinct. We do not promote m to a selection criterion—it would be a different normative choice (maximum joint construction rather than minimum total cost)—but report the alignment as a corroboration that, on this corpus, minimizing aggregate inferential effort tends to coincide with maximizing joint inferential construction, with Lia Lee identifying the kind of case where a mediator would have to choose explicitly. The mass profile $\big(m(h), a(h)\big)$ used here is heuristic; Section 6 derives its canonical counterpart—an axiomatically unique support geometry—and revisits the equity question with it.

5.5 Baselines and ablation: what the symbolic stages buy

We compare the cascade against a battery of baselines, from the naivest geometric fuser to the cascade with stages removed, and isolate the contribution of each stage. Throughout, "precision" is measured against the cascade-validated relation F; we return to the limits of this reference at the end.

Geometric baselines all fail. Table 4 reports three purely embedding-based methods: pairwise cosine thresholding, single-linkage graph clustering (connected components of the cosine graph—the same graph PN builds, but with cosine edges instead of NLI-validated edges), and spherical k-means clustering ($K = 200$); in each, cross-actor pairs accepted (or grouped) are taken as predicted convergences. None exceeds 7.2% precision against F. Single-linkage collapses into a single giant component, asserting essentially all cross-actor pairs ($\approx 0\%$ precision); cosine thresholding and k-means do little better. The reason is structural: embedding proximity captures *topic*, and is blind to the *polarity* that separates "same topic, convergent" from "same topic, opposed." No geometric method recovers the validated convergences.

Table 4: Geometric baselines—precision against F (% of predicted cross-actor convergences that are true cascade fusions). $|F|$ is the number of true fusions.

Case	$\lvert F\rvert$	Cosine-only (τ=0.55)	Single-linkage	k-means (K=200)
GERD	102	0.2%	$\approx 0\%$	0.8%
Bipolar I	38	0.1%	$\approx 0\%$	0.0%
Lia Lee	211	1.2%	$\approx 0\%$	7.2%
Selene	218	1.2%	$\approx 0\%$	3.7%

Stage ablation: NLI and role-inversion. The symbolic stages are what supply precision. Table 5 traces, from the pipeline's run logs, how each cascade stage reduces the candidate set. The cosine pre-filter removes the bulk of pairs cheaply (14–81%); bidirectional NLI removes the vast majority of the remainder (this is the *entailment-only* baseline—the cascade without role-inversion); and role-inversion then rejects a further 14–22% of the pairs NLI had accepted (19/121,

11/49, 35/246, 58/276)—pairs NLI judged equivalent but whose agency, causality, or polarity does not survive verification.

Table 5: Cascade stage ablation (pipeline run artifacts): pairs surviving each stage. "Pass NLI" is the entailment-only baseline (bidirectional NLI + contradiction veto, before role-inversion); the last column is the full cascade and equals $|F|$ of Table 4.

Case	Total pairs	After cosine pre-filter	Pass NLI (entailment-only)	After role-inversion
GERD	754,005	345,202	121	102
Bipolar I	252,003	216,111	49	38
Lia Lee	252,004	112,855	246	211
Selene	753,000	143,793	276	218

Each stage contributes and none is decorative: the neural pre-filter buys efficiency, NLI supplies most of the precision the geometric baselines lack, and role-inversion removes the residual false fusions that bidirectional entailment alone cannot catch. This is the sense in which the neuro-symbolic composition turns a cheap candidate set into a high-precision, auditable certificate.

The limit of these comparisons, and why no LLM can be the reference. Precision above is measured against F, the cascade's own output, so these results show that weaker geometric methods do not *recover* F, not that F is the ground truth. It is tempting to add an LLM-as-judge baseline, but it cannot serve as an independent reference: PN's own cascade *already includes* an LLM stage—the role-inversion verification—so an LLM re-judgment is a variant of a component of PN rather than an external arbiter, and would only check internal consistency. PN's notion of convergence is, by construction, model-relative (it is what the embedding, NLI, and LLM stages jointly certify). The only genuinely external validation of convergence *quality* is therefore human expert adjudication, which remains the key open item (Section 7.3).

5.6 Does convergence increase with depth? (semantic-collapse test)

A natural objection is that, as inferential trees deepen, consequences become more abstract ("stability", "efficiency", "risk reduction"), so that incompatible positions would converge *automatically* at depth, inflating HUBs with false positives—a "semantic collapse" in which depth $\uparrow \Rightarrow$ convergence $\uparrow$. We test this directly. For every node we record its tree depth, its participation in any cross-actor fusion, and its mean cosine similarity to nodes of the *other* actors (a geometric proxy for cross-actor convergence).

The effect does not appear. Across all four cases the Spearman correlation between node depth and cross-actor fusion rate is negligible (GERD 0.05, Jamison 0.04, Lia Lee -0.01, Selene 0.01), and the correlation between depth and mean cross-actor cosine is likewise near zero (GERD -0.06, Jamison 0.12, Lia Lee -0.06, Selene 0.00). The per-level mean cross-actor cosine is essentially *flat* with depth (e.g. GERD remains near 0.34 at every level from L_2 to L_8). Deeper nodes are neither more likely to fuse nor more similar across actors. We attribute this to two design elements that keep nodes anchored rather than drifting to generic abstractions: the conflict-relevance admission gate, which filters consequences that lose contact with Π, and the bicomposite weight, whose identity-drift term penalizes departure from the actor's root. The convergences PN reports are therefore not an artifact of depth-driven abstraction.[3]

[3]Computed in the released artifact, `semantic_collapse.py`.

6 A Canonical Support Geometry for the Certificate

The certificate F answers *whether* two propositions are equivalent under the cascade. It does not, by itself, answer a question every mediator asks of a detected convergence: *who builds it?* A HUB certified by pairwise NLI decisions may rest almost entirely on one agent's inferential material while the other agents merely countersign it with a handful of nodes. Section 5.4 introduced an inferential-mass profile to probe this; that construction, however, was heuristic. In this section we show that the objects the framework already constructs—the actor trees and the certified relation F—induce a *canonical* answer: a support geometry on the standard simplex, unique under four explicit axioms, in which every node, and in particular every HUB, occupies a position with an exact meaning. We then use this geometry to make the normative analysis of Section 3.3.1 two-dimensional: alongside *who pays* (TTF), the certificate now states *who builds*.

6.1 Support regions and the uniqueness theorem

Definition 9 (Support region axioms)**.** Let $V = \bigcup_i V_{a_i}$ with the parent–child relation of the actor trees, and let F be the certified fusion relation. For $x \in V$, write $\mathrm{sub}(x)$ for the set of descendants of x (including x). A *support-region operator* assigns to each x a set $R(x) \subseteq V$ such that:

A1 (inferential reflexivity) $\mathrm{sub}(x) \subseteq R(x)$: an agent's deployed consequences support the proposition they descend from.

A2 (substitution under certified equivalence) if $u \in R(x)$ and $(u, v) \in F$, then $\mathrm{sub}(v) \subseteq R(x)$: material certified biconditionally equivalent to part of x's support is itself support of x.

A3 (minimality) $R(x)$ is the smallest set satisfying A1–A2: nothing enters the region without a certified reason.

Proposition 1 (Existence and uniqueness)**.** For every x there is exactly one operator satisfying A1–A3: $R(x)$ is the least fixed point of the monotone map $S \mapsto \mathrm{sub}(x) \cup \bigcup\{\mathrm{sub}(v) : (u, v) \in F,\ u \in S\}$ on the finite lattice 2^V, which exists and is unique by the Knaster–Tarski theorem [14, 26]. Note that A2 iterates by construction: imported material can itself import. Truncating the iteration at any finite stage satisfies A1–A2 but violates A3, so the fixed point—not any one-step approximation—is the canonical region.

Definition 10 (Barycentric support coordinates)**.** With k actors, the *support coordinates* of x are $\lambda_i(x) = |R(x) \cap V_{a_i}| \,/\, |R(x)|$, a point of the standard simplex Δ^{k-1}. A node with $\lambda = (1, 0, \dots)$ is supported purely by its own actor; interior coordinates indicate certified cross-actor support.

One degree of freedom remains and we declare it rather than hide it: A1–A3 fix the *set* $R(x)$; the fourth axiom fixes the measure that λ places on it:

A4 (equal nodal weight) λ counts each proposition of $R(x)$ with equal weight — an insufficient-reason convention.

We probed this choice by re-computing λ with each node weighted by its descendant count: the per-HUB balance rankings correlate at Spearman $\rho = 0.79$–1.00 across the four cases, and every selection reported below (most-equitable HUB, and the divergence from the utilitarian optimum) is *identical* under both measures. The conclusions of this section do not depend on A4.

Two further properties situate the geometry. First, it is fully deterministic and parameter-free: position is computed from the tree topology and the certified pairs alone—no embedding,

no similarity score, no randomness—so two auditors with the same artifact reproduce the same point exactly. Second, it is certificate-relative in the sense of Section 7.3: λ inherits the backend-dependence of F, and all claims below are qualitative.

6.2 HUBs are points, and the support demography

Across the four cases, the members of any one strict-HUB closure collapse to a *single* point of the simplex: the median pairwise λ-distance within a closure is 0.000 in all four corpora. This is the geometric face of what the closure already is logically—mutually certified material shares its support region—and it licenses speaking of *the* position of a HUB. Distinct HUBs, by contrast, occupy distinct positions (all 3 closures in GERD; 11 closures on 6 positions in Bipolar I; 37 closures on 12 positions in Lia Lee, a meso-scale grouping consistent with the dominant convergence classes of Section 5.3). The exception is instructive: in Selene, whose fusion graph concentrates into a single dense component, the two closures share one position—when F is strongly entangled, the geometry loses resolution, a case-level signature rather than a defect.

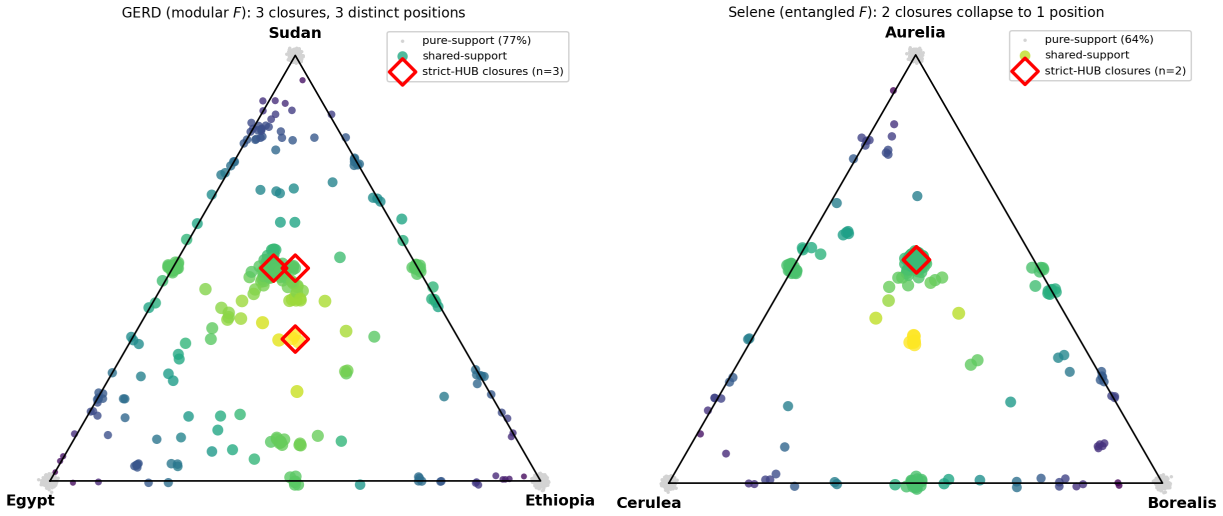

Figure 7: Support simplex for the two three-actor cases. Gray: pure-support nodes. Colored: nodes with certified cross-actor support (size and color track the mixing $1 - \max_i \lambda_i$). Red diamonds: strict-HUB closures, each collapsing to a single point (median within-closure λ-distance 0.000 in all four corpora). Left, GERD: a modular fusion graph yields three HUBs at three distinct positions. Right, Selene: an entangled fusion graph collapses both closures onto one position—the resolution of the geometry is itself a case-level signature. Position is computed from tree topology and certified pairs only.

The corpus demography is stable across the four cases: 62–84% of nodes are pure (they sit at their actor's vertex); 6–23% are F-endpoints; and a further 10–15% are *carriers*—nodes with no certified fusion of their own whose region imports shared material from their descendants. Carriers are systematically shallower than both other populations (mean depth 3.3–3.7 vs. ≈ 5): fusion happens deep in the trees, and shared support propagates upward to mid-level strategic propositions. We note for completeness that we also tested, under a pre-registered protocol, whether the mixed interior of the simplex *predicts* convergences the cascade had not certified (by submitting interior cross-actor

pairs to the cascade against random controls): it does not (0/300 vs. 0/300). The geometry is a canonical representation of the certificate, not an additional detector, and we report it as such.

6.3 Who builds vs. who pays: a two-dimensional equity reading

For a HUB h with position $\lambda(h)$ and cost vector $(\mathrm{TTF}(a_i, h))_i$, define two balance scores in $[0, 1]$: construction balance $\beta_\lambda(h) = \min_i \lambda_i(h) / \max_i \lambda_i(h)$ and cost balance $\beta_{\mathrm{TTF}}(h) = \min_i \mathrm{TTF}(a_i, h) / \max_i \mathrm{TTF}(a_i, h)$. The two dissociate empirically: in GERD, one HUB is built 50% from Sudan's material yet is cheapest for Ethiopia to reach; in Bipolar I, most HUBs are built 74–94% from the manic state's material—convergences the depressive state co-signs but does not construct—which sharpens, in canonical form, the asymmetry the mass profile of Section 5.4 could only suggest heuristically.

Table 6: The price of equity. For each case: the HUB selected by the utilitarian rule of Definition 6 vs. the HUB maximizing the equity score $\beta_\lambda \cdot \beta_{\mathrm{TTF}}$. The utilitarian optimum is never the most equitable convergence available; equity is purchased with additional total friction.

	Utilitarian $\arg\min \mathrm{TTF}_{\mathrm{sum}}$			Most equitable ($\beta_\lambda \beta_{\mathrm{TTF}}$)		
Case	$\mathrm{TTF}_{\mathrm{sum}}$	β_λ	equity rank	$\mathrm{TTF}_{\mathrm{sum}}$	β_λ	cost premium
GERD	17.6	0.41	3/3	21.4	1.00	+21%
Bipolar I	8.3	0.33	8/11	12.7	1.00	+53%
Lia Lee	7.2	0.75	24/37	13.0	1.00	+80%
Selene	17.1	0.46	2/2	18.0	0.46	+5%

Table 6 extends the normative analysis of Section 3.3.1 with the dimension it lacked. In all four cases the utilitarian-selected HUB is *not* the most equitable available convergence—in GERD it is the least equitable of the three, and the construction-balanced alternative ($\beta_\lambda = 1.00$, the African-Union mediation HUB) costs 21% more total friction. This does not overturn the selection rule; it completes its disclosure. PN's commitment to total efficiency over distributive equity was already stated as a debatable normative choice; the support geometry now prices that choice case by case, and a mediator who prefers the egalitarian reading can select from the same certified menu with both columns in view. Consistent with the framework's stance throughout, the geometry returns an explanation—who builds, who pays, at what premium—not a verdict.

7 Analysis and Discussion

7.1 Related work and positioning

Neuro-symbolic frameworks differ chiefly in how the neural and symbolic components are coupled. Tightly integrated approaches—Logic Tensor Networks [23], DeepProbLog [16]—embed logical or probabilistic constraints into neural training, letting symbolic structure be learned or relaxed by gradient descent. PN takes the opposite stance within the compositional landscape of Garcez and Lamb [8]: the neural layer is a candidate generator, and the symbolic layer is an auditable certificate the neural layer cannot modify. This inviolability is what lets every convergence be traced to its per-pair NLI decisions and per-actor paths, and is the property that situates PN in *explainable* neuro-symbolic AI. Adjacent symbolic traditions—abstract argumentation [5]—model defeasibility among arguments but presuppose a unified evaluator and do not address structural convergence across actor-relative trees; and topological analysis of representation geometry [21] is used here only

as post-hoc visualization, never as a constituent of detection. To our knowledge, no prior neuro-symbolic framework addresses multi-actor convergence detection from incompatible premises while emitting a structured explanation as its output.

7.2 Relation to social choice and bargaining

Arrow's impossibility theorem [2] concerns aggregation over a fixed exogenous alternative set under declared ordinal preferences. PN sits outside its premises: candidate HUBs are constructed *endogenously* from inferential trees, and actors supply inferential structures T_{a_i}, not orderings. PN also differs from AGM belief revision [1]: it operates on the consequence structures derived from premises, not on the premises themselves. Within the structural domain induced by the trees, PN's selection rule is the *utilitarian* bargaining solution. We note that its Pareto efficiency (Proposition 1) follows by an elementary argument and is not, in itself, the contribution; the substantive content is the *identification* of PN's lexicographic rule with the utilitarian solution and its empirical divergence from Nash (the unequal-split case of GERD, Section 3.3.1). Making this normative commitment explicit—total efficiency over distributive equity—is part of the explanation PN provides.

7.3 Limitations

We separate what is demonstrated from what is, at this stage, illustrated.

Dependence on the inference model (the central caveat). PN's entire search space is the set of LLM-generated inferential trees, and its notion of equivalence is operationalized by the NLI model: $F(P, Q)$ holds when the cross-encoder so judges. PN therefore detects convergence *as construed by these models*, not convergence in a model-independent sense. The fictional control (Section 5.1) shows the framework does not merely recall known outcomes—an invented case has none—but it does not establish independence from general training-distribution bias, nor that the trees are complete or unbiased. This is the framework's principal epistemological limit. Two consequences follow. First, the NLI signal is reliable for *entailment* (on which fusion depends) but its *contradiction* label, trained on single-premise inference, conflates logical opposition with mere alternativity in some constructions; claims resting on the contradiction label (the Dung analysis) should be read with this caveat. Second, an independent, human validation of HUB quality is required and remains the key open item (below).

The generative dependence is removable (a spectrum, not a fixed cost). The caveat above is a property of the *automated instantiation*, not of the framework. Tree generation and equivalence certification are two separable neural touchpoints, and only the first injects training-distribution bias into the search space. Replacing the LLM generator with *human-authored* inferential trees—supplied by the parties themselves, clinicians, or court-appointed experts—removes that bias entirely while leaving the symbolic machinery (region closure, λ, TTF, HUB extraction) untouched, since none of it depends on how the nodes were produced. This defines a spectrum of instantiation: fully automated (maximal scale, the generative caveat in force but *exposed*, never hidden); human-authored trees with NLI certification (zero generative bias, the equivalence step a narrower, auditable, swappable measurement); or human-authored trees with symbolic fusion (no neural component at all, maximal auditability, minimal scale). PN therefore does not eliminate the neural black box so much as *corner* it to declared points, expose every place it is consulted as an artifact, and make it removable at the analyst's discretion—the property that separates an auditable derivation from a black-box verdict.

Backend specificity, structural invariance. A specific instantiation of the previous caveat: the set of certified F-edges, the resulting clique cores, and any per-HUB statistics are conditional on the particular NLI cross-encoder used (DeBERTa-v3-base in this work). To probe this dependence empirically, we replicated the cascade with an independently trained RoBERTa-base [15] cross-encoder under matched operating thresholds; 13%–56% of strict-HUB core pairs were re-certified across the four cases. The *quantitative* HUB inventory is therefore calibration-relative. The *qualitative* structural properties of F, by contrast, are preserved across both backends: F is sparse (density < 0.10), exhibits low triadic clustering (< 0.15), has bounded k-core depth (≤ 4), and is predominantly arboreal ($\geq 78\%$ of edges in a spanning forest). This dependence is intrinsic to any neuro-symbolic framework built on third-party inference modules: the framework's precision tracks the precision of its components. Conversely, progress in cross-encoder textual inference—through scale, domain-adapted corpora, or compositional entailment models—translates directly into improved reliability of PN without architectural change, since the downstream graph construction, HUB extraction, TTF computation, and bargaining selection are invariant to the choice of NLI head.

Aggregation-criterion robustness (AVG vs. MIN). The corpus is certified at the AVG operating point of Section 3.2. Re-certifying every pair from the cached NLI evaluations under the strictly stronger MIN aggregate—each direction individually $\geq \tau_{ent}$, contradiction veto and role-inversion unchanged—retains 74–86% of F (GERD 77/102, Jamison 31/38, Lia Lee 182/211, Selene 162/218) and the bulk of the unique strict-HUB closures (2/3, 10/11, 35/37, 1/2). The utilitarian selection is essentially invariant: the selected HUB survives with identical closure and text in GERD ($\text{TTF}_{\text{sum}} = 17.61$) and Jamison (8.26); in Selene the selected closure survives with a different representative core ($17.11 \rightarrow 17.96$); only in Lia Lee does the cost-minimal HUB not survive, the selection shifting to an adjacent HUB of the same monitoring convergence class ($7.24 \rightarrow 7.94$). One structural casualty is worth flagging: GERD's Nash-preferred alternative ($\text{TTF}_{\text{sum}} = 20.36$, Section 3.3.1) does not survive MIN, so the utilitarian-vs-Nash contrast is a property of the AVG operating point. The check is deterministic and model-free (computed from the released per-pair artifacts in `min_recert_check.py`).

HUB-quality validation by domain experts is open (the most important next step). PN's central claim is that detected HUBs are valid convergences, and the structural guarantees offered in this paper (pairwise cascade certificates, Pareto efficiency, no argumentative dominance, absence of depth-driven collapse in Section 5.6, robustness of the ranking to the bicomposite-weight variant in Table 7) constrain but do not replace domain-specific assessment. We explicitly do *not* claim that the HUBs we report are substantively correct beyond their cascade certificate; that judgment requires *domain experts* (hydropolitical specialists for GERD, clinicians for Jamison, medical anthropologists for Lia Lee) reading each detected HUB and ruling whether the convergence is sound in its native discourse. An earlier version of this framework received external recognition at the 2026 GCSP Prize for Transformative Futures in Peace and Security (Second Place, 337 submissions, 95 countries, jury including ETH Zurich, Oxford, and the Swiss Federal Administration), but a research-grade prize is not peer-reviewed domain-expert validation, and we draw the distinction explicitly: the GCSP recognition speaks to the framework's relevance and presentation; substantive HUB-quality validation remains the principal piece of empirical work that this paper does not yet carry. Until that validation is performed, the framework should be read as a structured convergence-*discovery* procedure with a transparent certificate—not as a validated theory of latent convergence.

Heuristic, not axiomatic, metric. The bicomposite weight (equal, additive identity-drift and conflict-relevance terms) is a principled design choice —both terms non-negative and interpretable, ensuring Dijkstra correctness—but it is not derived from axioms, and alternatives (unequal weights, product forms, learned metrics) are not excluded. To probe whether this choice is load-bearing, we recomputed $\mathrm{TTF}_{\mathrm{sum}}$ for every HUB under four alternative edge weights: multiplicative $(1-\mathrm{sim}_R)(1-\mathrm{sim}_\Pi)$, logarithmic $-\log(\mathrm{sim}_R \cdot \mathrm{sim}_\Pi)$, identity-drift only $(1-\mathrm{sim}_R)$, and conflict-relevance only $(1-\mathrm{sim}_\Pi)$. Table 7 reports the Spearman rank correlation of the HUB ranking against the additive baseline, computed on the canonical corpus (unique strict closures, each represented by its cost-minimal instance). On the two cases with a rankable inventory ($H \geq 11$) the correlation is ≥ 0.94 across all four variants, and the utilitarian-selected top-1 HUB matches the additive baseline in every (case, variant) cell, 16/16. The HUB ranking is therefore robust to the specific functional form of the bicomposite weight rather than driven by it; the additive form is one of several admissible choices that yield substantively the same selection.

Table 7: Robustness of HUB ranking to bicomposite-weight variant. Spearman ρ of $\mathrm{TTF}_{\mathrm{sum}}$ ranking against the additive baseline, and whether the top-1 (utilitarian-selected) HUB matches. H is the number of unique strict-HUB closures of Section 5. GERD ($H = 3$) and Selene ($H = 2$) admit few HUBs and are shown for completeness. Computed on the canonical corpus from the released artifact (`ablation_ttf_canonical.py`); the additive variant reproduces the per-HUB $\mathrm{TTF}_{\mathrm{sum}}$ of the case artifacts exactly.

		Spearman ρ vs additive				Top-1 matches			
Case	H	mul	log	id_R	id_Π	mul	log	id_R	id_Π
GERD	3	1.00	1.00	1.00	1.00	✓	✓	✓	✓
Jamison	11	0.95	0.95	0.97	0.99	✓	✓	✓	✓
Lia Lee	37	0.94	0.95	0.96	0.97	✓	✓	✓	✓
Selene	2	1.00	1.00	1.00	1.00	✓	✓	✓	✓

Validation-set size. Threshold calibration rests on 53 labeled pairs and NLI model selection on 21 cases; these are calibration sets, not benchmarks, and are small by contemporary standards. Larger labeled benchmarks would strengthen the threshold and model choices.

Other limits. Tree quality reflects inference-engine biases (mitigated by ensemble expansion); the bicomposite weight depends on the relevance of Π; where rejection is axiomatic no HUB exists and the algorithm reports this as a valid diagnostic; embedding anisotropy could in principle compress the dynamic range of w along long accumulated paths, although the measured cross-actor cosine distributions in our corpus do not show pathological compression; and although the embedding pre-filter empirically removes 14–81% of pairs, the pairwise NLI stage remains the computational bottleneck of the cascade.

7.4 Why the NLI cascade and not geometric clustering

A natural question is whether the three NLI-based cascade stages could be replaced or reduced to a purely geometric procedure: clustering on the embedding manifold, distance thresholds, density estimation, or convex-region membership. We rule this out on structural grounds. The cosine-similarity distribution of the corpus is organized against geometric methods. Across the four cases,

the modularity gap (mean within-actor cosine minus mean across-actor cosine) ranges only from 0.065 to 0.085 (Table 8). The within-actor and across-actor cosine distributions overlap massively (Figure 8); actors do not occupy disjoint regions of the embedding manifold.

Yet the relation F certified by the cascade rejects approximately 97% of cross-actor pairs with cosine ≥ 0.6 (Table 9). The discriminative structure that separates actors logically is therefore *not* aligned with cosine similarity: it lives along a direction the NLI head detects but that distance-based methods cannot recover. The cascade is not a refinement of geometric clustering; it operates on an independent semantic axis. A clustering rule would conflate "same topic, compatible" with "same topic, opposed," because both categories sit in the same region of the manifold.

The intrinsic dimensionality of the manifold further constrains what geometric methods could achieve. Across the four cases, 90–95% of the variance is captured by only 80–103 principal components out of 768 nominal dimensions, with anisotropy ratios (dominant eigenvalue divided by mean eigenvalue) of 14–94. Cross-actor distances are therefore distributed in a high-dimensional regime in which density estimation, convex-region membership, and other geometric heuristics lose discriminative power.

Table 8: Cosine-similarity structure of the corpus. $\bar{c}_{\text{within}}$ and $\bar{c}_{\text{across}}$ are mean within-actor and mean cross-actor cosine; the modularity gap $\Delta\bar{c}$ is their difference. r counts principal components capturing $\geq 90\%$ of the variance; α is the anisotropy ratio (top eigenvalue over mean eigenvalue).

Case	$\bar{c}_{\text{within}}$	$\bar{c}_{\text{across}}$	$\Delta\bar{c}$	r at 90%	α
GERD	0.39	0.32	0.07	103	14
Jamison	0.34	0.27	0.07	80	94
Lia Lee	0.31	0.22	0.09	98	23
Selene	0.42	0.34	0.08	93	17

Table 9: Discriminative gap between cosine and cascade certification. Among cross-actor pairs with cosine ≥ 0.6 (high geometric proximity), the fraction certified by F is small in every case: cosine similarity does not predict cascade-certified equivalence.

Case	#pairs cos ≥ 0.6	#certified by F	% certified
GERD	612	19	3.1%
Jamison	1487	44	3.0%
Lia Lee	3091	96	3.1%
Selene	221	8	3.6%

We conclude that the NLI cascade is not a refinement of, but an alternative to, geometric clustering: it certifies pairs on a logical axis (mutual entailment under the cascade) that cuts across the geometric one, and produces a relation whose triadic closure cannot be recovered from any partition of the cosine grid. The triangle closure of F is therefore not a property that any geometric partition can recover by locality alone; transitive closure of the certified relation is constructed by Union–Find over verified pairs, not extracted from spatial neighborhoods.

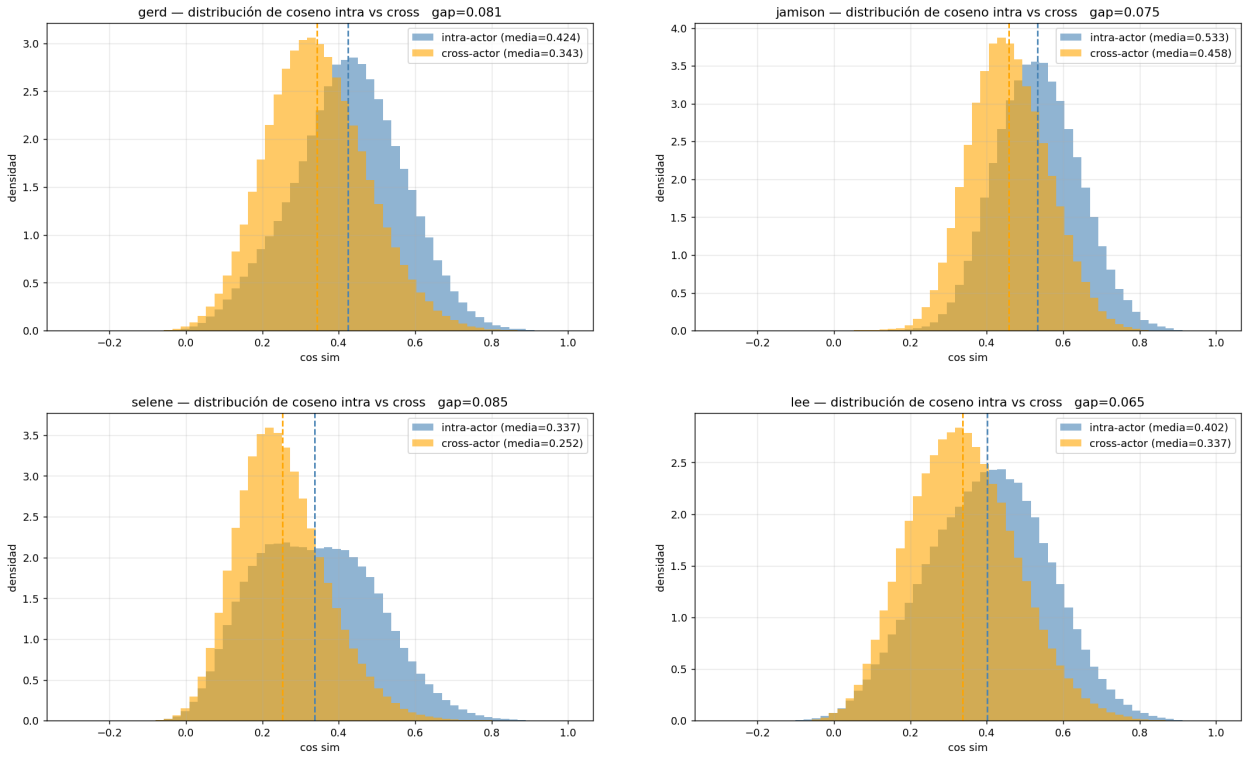

Figure 8: Within-actor (orange) and cross-actor (blue) cosine similarity distributions. The distributions overlap massively in every case: actors do not occupy disjoint regions of the embedding manifold, so geometric clustering cannot reproduce the cascade's certification.

8 Conclusions and Future Work

We presented the Principle of Naturality, an explainable neuro-symbolic framework that detects latent inferential convergence among agents with mutually exclusive premises and returns, as its primary output, a structured and auditable explanation of that convergence: the convergence point (HUB), the per-agent causal paths that reach it, the per-pair certificates that license it, and the normative justification of its selection as a Pareto-efficient utilitarian bargaining solution. The neural layer generates candidates; the symbolic layer issues a certificate it cannot overrule; a post-hoc topological layer renders the shape of the disagreement. The certificate, moreover, is not only a list of validated pairs: under four explicit axioms it induces a unique support geometry (Section 6) in which every HUB occupies a single barycentric point and the equity of every convergence—who builds it, who pays to reach it, and the price of preferring the balanced alternative—is itself part of the auditable output.

The empirical evaluation supports both hypotheses across four heterogeneous configurations. The fully fictional Selene control isolates the framework's behavior from language-model memorization: PN emits no convergence when the implication space is closed by construction, and finds convergences when a reachable common point exists. GERD yields the strongest topological signal of the corpus, and Sudan emerges as the structural pivot of the utilitarian-selected HUB. The selected HUB is Pareto-efficient in 4/4 cases, and its divergence from Nash in GERD makes PN's distributive stance explicit and inspectable—which is what an explanation, as opposed to a verdict, must do.

Why the auditable form matters. In the settings PN targets—diplomacy, clinical ethics, dispute mediation—a convergence a party cannot inspect is a convergence a party cannot trust, and an opaque agreement score, however accurate, gives a mediator no defensible basis to act. PN's contribution is therefore not only that it *detects* latent agreement but that it returns it in a contestable form: every HUB decomposes into per-pair certificates and per-actor paths a domain expert can accept or reject one at a time. This substitutes scrutiny of an explicit derivation for trust in a black box—the property that, more than any single number, we take to be the result.

Future work. PN's machinery is, in principle, a general instrument for characterizing and comparing how different agents—or different cognitive architectures—organize the same problem. A companion study applies the same pipeline to compare axiomatically defined cognitive architectures (e.g. autistic and neurotypical cognition), using the topological layer to visualize the *shape* of each architecture's reasoning and the cascade to locate where two architectures can and cannot meet. Beyond conflict resolution, the same structured explanation could support adjudication, reasoning analysis, and clinical or mental-health modeling. We draw no clinical conclusions here; such applications require independent domain validation and human expert supervision.

Ethical considerations. PN identifies regions of cross-actor inferential compatibility, not objectively shared truths. A detected HUB is a starting point for negotiation, not an externally imposed solution, and could be misused if presented as the latter; we recommend operation under human expert supervision in sensitive domains. The case studies use publicly available material only.

Data and code availability. The certified graphs of the four cases, all analysis scripts, and the JSON artifacts behind every reported number are archived at Zenodo (DOI: 10.5281/zenodo.18620115); Appendix G lists the validated default configuration, and every certification stage declares its backend model and version.

A Mathematical Notation

Principal symbols: $T_{a_i} = (V_{a_i}, E_{a_i})$ actor a_i's inferential tree; R_i its root (declared position); Π the shared problem statement; $\varphi : U \to \mathbb{R}^{768}$ the embedding map; sim cosine similarity; $\tilde{P}$ the pragmatic-normalization operator; F the fusion relation; core_{strict}, core_{lax} the validated cores; $w(u, v)$ the bicomposite edge weight; TTF the Total Topological Friction with aggregates TTF_{sum}, $\text{TTF}_{\max}$; H_0, H_1 persistent-homology dimensions.

B Causal Inference Protocol

This protocol defines the restrictions for LLM-driven causal tree construction.

R1 Direct causal implication. Each edge (P, Q) must represent a direct causal implication—"if P then Q"—defensible from the actor's perspective without implicit intermediate steps.

R2 Perspective preservation. Consequences derive from the actor's perspective; the engine introduces no external information and does not correct the actor's biases.

R3 Propositional atomicity. Each node is a single evaluable proposition admitting a yes/no answer to "is it true that [proposition]?".

R4 Cycle prohibition. No consequence may imply a previous premise.

R5 Termination. Expansion stops at a terminal state, a redundant consequence, depth $d_{\max}$, or node budget $n_{\max}$.

R6 Actor isolation (blind expansion). The engine generates one actor's consequences without visibility of other actors' positions.

R6.5 Third-person normalization. All generated propositions refer to the actor by their alias α_i in third person, so the operator $\tilde{P}$ (Section 3.2) finds its substitution target.

C Solution Synthesis Protocol

Solution synthesis (Phase 5) is optional and non-algorithmic: PN's formal guarantee covers *detection* (Phases 1–4). Synthesis (1) selects the optimal HUB; (2) extracts the common elements of the convergence paths; (3) formulates a proposal in accessible language; (4) checks the proposal against each actor's root premises. It introduces subjectivity and must be treated as assistance to a human mediator, not an automatic solution.

D Algorithm Pseudocode

```
ALGORITHM PrincipleOfNaturality(actors, premises, Pi, params)
  for each (actor, premise): trees += BuildCausalTree(actor, premise, Pi, params)
  G            <- MergeGraphs(trees)                      # Phase 2a
  F            <- DetectFusions(G, params)                # Phase 2b (cascade)
  groupings    <- UnionFind(F)                            # transitive closure
  hubs         <- FindHubs(G, actors, params)             # Phase 4
  return hubs ordered by (tau, TTF_sum, TTF_max)

FUNCTION BuildCausalTree(actor, root, Pi, d_max, n_max, tau_adm)
  pq <- {(root, level=0, priority=0)}
  while pq not empty and nodes_created < n_max:
    (node, level) <- pq.extract_min()
    if level >= d_max: continue
    for c in LLM_InferConsequences(actor, node, ancestors):
      sim_Pi <- sim(phi(c), phi(Pi))
      if sim_Pi < tau_adm: continue                       # admission gate
      W_accum <- W(node) + (1 - sim_root) + (1 - sim_Pi)  # bicomposite
      if not redundant: pq.insert(c, priority=W_accum)
  return T
```

In all reported runs $\tau_{\text{adm}} = 0.15$, and the cross-branch redundancy filter discards a candidate whose actor-normalized cosine to an existing tree node exceeds 0.90.

E NLI Model Validation

On a 21-case English validation set, DeBERTa [10] and MiniLM [27] both fail only Case 2 (an asymmetric kinship relation, a known NLI limitation absent from causal-tree propositions); MiniLM's

additional failures concentrate in problematic and positive-bias categories, motivating DeBERTa (`cross-encoder/nli-deberta-v3-base`) as the primary model.

F Threshold Empirical Validation

On 53 labeled proposition pairs, the thresholds $\tau_{ent} = 0.55$ and $\tau_{contr} = 0.30$ produce correct fusion/contradiction classifications; $\tau_{ent} = 0.55$ lies within the maximum-F_1 zone for both models and variations of ± 0.10 do not alter the results. DeBERTa attains precision 1.000 in fusion detection under the bidirectional criterion of Section 3.2.

G Default Configuration

Parameter	Value
Embedding model	`paraphrase-multilingual-mpnet-base-v2`
NLI model	`cross-encoder/nli-deberta-v3-base`
NLI aggregation	AVG over both directions + contradiction veto (MIN as robustness re-certification)
LLM (expansion)	`claude-sonnet`, temperature 0
τ_{pre} (pre-filter)	0.35
τ_{ent} (fusion)	0.55
τ_{contr} (contradiction veto)	0.30
$d_{\max}$ (depth)	10
$n_{\max}$ (nodes/actor)	500
TDA permutations	500

References

[1] Carlos E. Alchourrón, Peter Gärdenfors, and David Makinson. On the logic of theory change: Partial meet contraction and revision functions. *Journal of Symbolic Logic*, 50(2):510–530, 1985.

[2] Kenneth J. Arrow. *Social Choice and Individual Values*. Yale University Press, 1951.

[3] Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. A large annotated corpus for learning natural language inference. In *Proceedings of EMNLP*, pages 632–642, 2015.

[4] Edsger W. Dijkstra. A note on two problems in connexion with graphs. *Numerische Mathematik*, 1(1):269–271, 1959.

[5] Phan Minh Dung. On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. *Artificial Intelligence*, 77(2):321–357, 1995.

[6] Anne Fadiman. *The Spirit Catches You and You Fall Down: A Hmong Child, Her American Doctors, and the Collision of Two Cultures*. Farrar, Straus and Giroux, 1997.

[7] Bernard A. Galler and Michael J. Fischer. An improved equivalence algorithm. *Communications of the ACM*, 7(5):301–303, 1964.

[8] Artur d'Avila Garcez and Luís C. Lamb. Neurosymbolic AI: The 3rd wave. *Artificial Intelligence Review*, 56(11):12387–12406, 2023.

[9] Frederick K. Goodwin and Kay R. Jamison. *Manic-Depressive Illness: Bipolar Disorders and Recurrent Depression*. Oxford University Press, 2 edition, 2007.

[10] Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. DeBERTa: Decoding-enhanced BERT with disentangled attention. In *International Conference on Learning Representations (ICLR)*, 2021.

[11] Kay R. Jamison. *An Unquiet Mind: A Memoir of Moods and Madness*. Alfred A. Knopf, 1995.

[12] Ehud Kalai. Proportional solutions to bargaining situations: Interpersonal utility comparisons. *Econometrica*, 45(7):1623–1630, 1977.

[13] Henry A. Kautz. The third AI summer: AAAI robert s. engelmore memorial lecture. *AI Magazine*, 43(1):105–125, 2022.

[14] Bronisław Knaster. Un théorème sur les fonctions d'ensembles. *Annales de la Société Polonaise de Mathématique*, 6:133–134, 1928.

[15] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

[16] Robin Manhaeve, Sebastijan Dumančić, Angelika Kimmig, Thomas Demeester, and Luc De Raedt. DeepProbLog: Neural probabilistic logic programming. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 31, 2018.

[17] Hugo Mercier and Dan Sperber. Why do humans reason? arguments for an argumentative theory. *Behavioral and Brain Sciences*, 34(2):57–74, 2011.

[18] John F. Nash. The bargaining problem. *Econometrica*, 18(2):155–162, 1950.

[19] Elinor Ostrom. *Governing the Commons: The Evolution of Institutions for Collective Action*. Cambridge University Press, 1990.

[20] Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using siamese BERT-networks. In *Proceedings of EMNLP-IJCNLP*, pages 3982–3992, 2019.

[21] Manish Saggar, Olaf Sporns, Javier Gonzalez-Castillo, Peter A. Bandettini, Gunnar Carlsson, Gary Glover, and Allan L. Reiss. Towards a new approach to reveal dynamical organization of the brain using topological data analysis. *Nature Communications*, 9:1399, 2018.

[22] Thomas C. Schelling. *The Strategy of Conflict*. Harvard University Press, 1960.

[23] Luciano Serafini and Artur d'Avila Garcez. Logic tensor networks: Deep learning and logical reasoning from data and knowledge, 2016. arXiv:1606.04422.

[24] Gurjeet Singh, Facundo Mémoli, and Gunnar Carlsson. Topological methods for the analysis of high dimensional data sets and 3D object recognition. In *Eurographics Symposium on Point-Based Graphics*, pages 91–100. The Eurographics Association, 2007.

[25] Robert E. Tarjan. Efficiency of a good but not linear set union algorithm. *Journal of the ACM*, 22(2):215–225, 1975.

[26] Alfred Tarski. A lattice-theoretical fixpoint theorem and its applications. *Pacific Journal of Mathematics*, 5(2):285–309, 1955.

[27] Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. MiniLM: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. In *Advances in Neural Information Processing Systems 33 (NeurIPS)*, 2020.

[28] Adina Williams, Nikita Nangia, and Samuel R. Bowman. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of NAACL-HLT*, pages 1112–1122, 2018.